Appendix A. Snapshot into the State of ML4H Model Evaluation

To get a snapshot of the current standards for model evaluation in machine learning for healthcare research, we manually reviewed all of the papers from the CHIL 2022 proceedings, the first 20 papers in the CHIL 2021 proceedings, and the first 20 papers that came up in the Radiology medical journal when searching for the keyword "machine learning" and filtering for papers from 2022 to 2023 (see README.md in https://github.com/acmi-lab/EvaluationOverTime). Out of 23 papers in the CHIL 2022 proceedings, 21 did not take time into account in their data split, and two were unclear about how they split data, but it is unlikely that they split by time. Out of the 20 papers reviewed at CHIL 2021, only one paper split by time. Out of the 20 papers reviewed from Radiology, 6 did not train or evaluate any machine learning models, but out of the remaining 14 papers, 13 did not take time into account in their data split, and one did not specify how data was split.

Appendix B. EMDOT Python Package

Figure 6 illustrates the workflow of the EMDOT Python package.

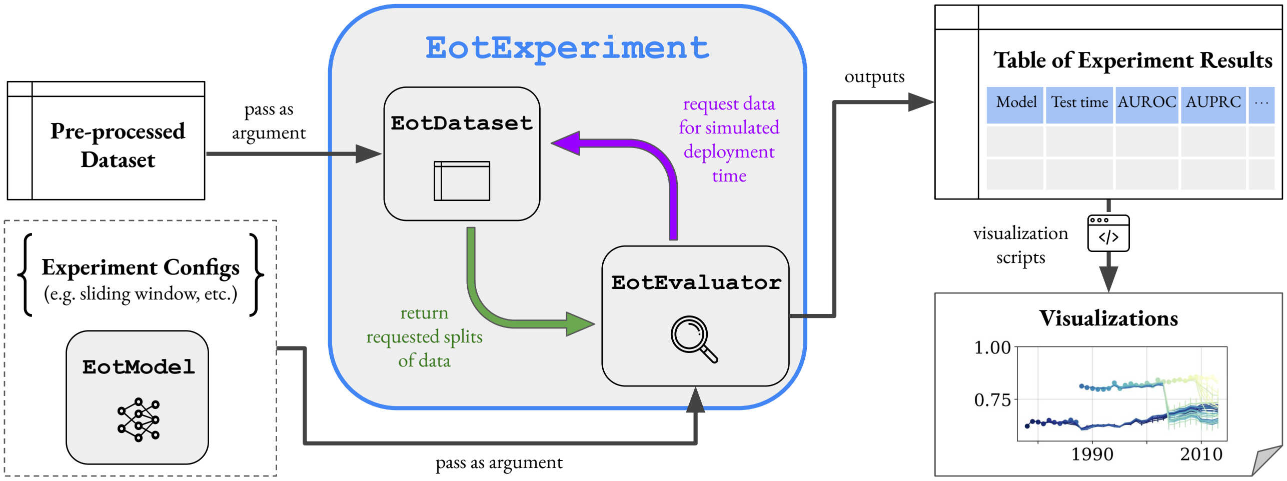


Figure 6: EMDOT Python package workflow diagram. The primary touchpoint of the EMDOT package is the `EotExperiment` object. Users provide a dataframe for their (mostly) preprocessed dataset (EMDOT takes care of normalization based on the relevant training set), their desired experiment configuration (e.g. sliding window), and model class (which should subclass the simple `EotModel` abstract class) in order to create an `EotExperiment` object. Running the `run_experiment()` function of the `EotExperiment` returns a dataframe of experiment results that can then be visualized. The diagram also provides insight into some of the internals of the `EotExperiment` object – there is an `EotDataset` object that handles data splits, and an `EotEvaluator` object that executes the main evaluation loop.

Appendix C. Additional SEER Data Details

The Surveillance, Epidemiology, and End Results (SEER) Program collects cancer incidence data from registries throughout the U.S. This data has been used to study survival in several forms of cancer (Choi et al., 2008; Fuller et al., 2007; Taioli et al., 2015; Hegselmann et al., 2018). Each case includes demographics, primary tumor site, tumor morphology, stage and diagnosis, first course of treatment, and survival outcomes (collected with follow-up) (National Cancer Institute, 2020). The performance over time is evaluated on a *yearly* basis. We use the November 2020 version of the SEER database with nine registries (SEER 9), which covers about 9.4% of the U.S. population. While there are SEER databases that aggregate over more registries and hence cover a greater proportion of the U.S. population, we choose SEER 9 due to the large time range it covers (1975–2018).

- Data access: After filling out a Data Use Agreement and Best Practices Agreement, individuals can easily request access to the SEER dataset.

- Cohort selection: Using the SEER*Stat software (Program, 2015), we define three cohorts of interest: (1) breast cancer, (2) colon cancer, and (3) lung cancer. We primarily follow the cohort selection procedure from Hegselmann et al. (2018), but we use SEER 9 instead of SEER 18, and use data from all available years instead of limiting to 2004–2009. Cohort selection diagrams are given in Figures 7, 8, and 9. If there are multiple samples per patient, we filter to the first entry per patient, which corresponds to when a patient first enters the dataset. This corresponds to a particular interpretation of the prediction: when a patient is first added to a cancer registry, given what we know about that patient, what is their estimated 5-year survival probability?

- Cohort characteristics: Summaries of the SEER (Breast), SEER (Colon), and SEER (Lung) cohort characteristics are in Tables 3, 4, and 5.

- Outcome definition: 5-year survival is defined by a confirmation that the patient is alive five years after the year of diagnosis.

- Features: We list the features used in the SEER breast, colon, and lung cancer datasets in Section C.2. For all datasets, we convert all categorical variables into dummy features, and apply standard scaling to numerical variables (subtract mean and divide by standard deviation).

- Missingness heat maps: are given in Figures 10, 11, 12, 13, 14, and 15.

C.1. Cohort Selection and Cohort Characteristics

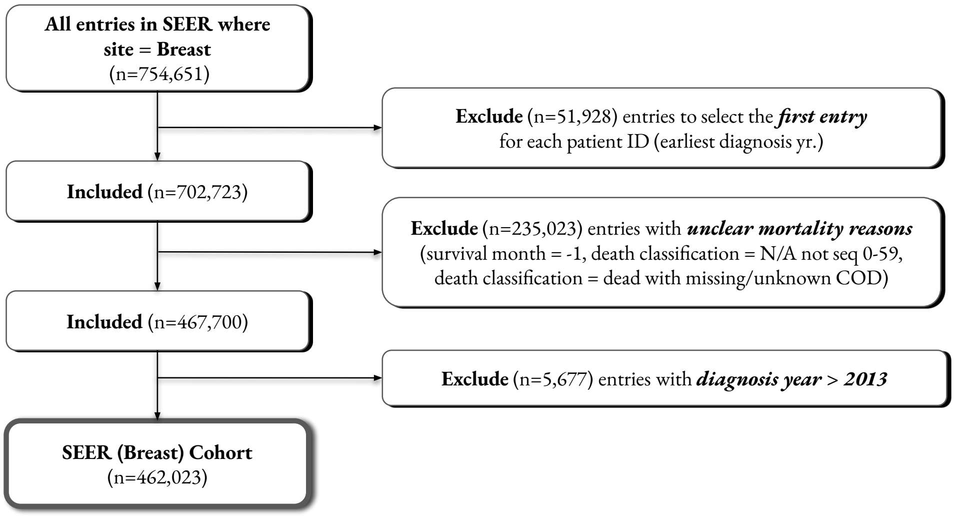

All entries in SEER where site = Breast (n=754,651)
Exclude (n=51,928) entries to select the first entry for each patient ID (earliest diagnosis yr.)
Included (n=702,723)
Exclude (n=235,023) entries with unclear mortality reasons (survival month = -1, death classification = N/A not seq 0-59, death classification = dead with missing/unknown COD)
Included (n=467,700)
Exclude (n=5,677) entries with diagnosis year > 2013
SEER (Breast) Cohort (n=462,023)

Figure 7: Cohort selection diagram - SEER (Breast)

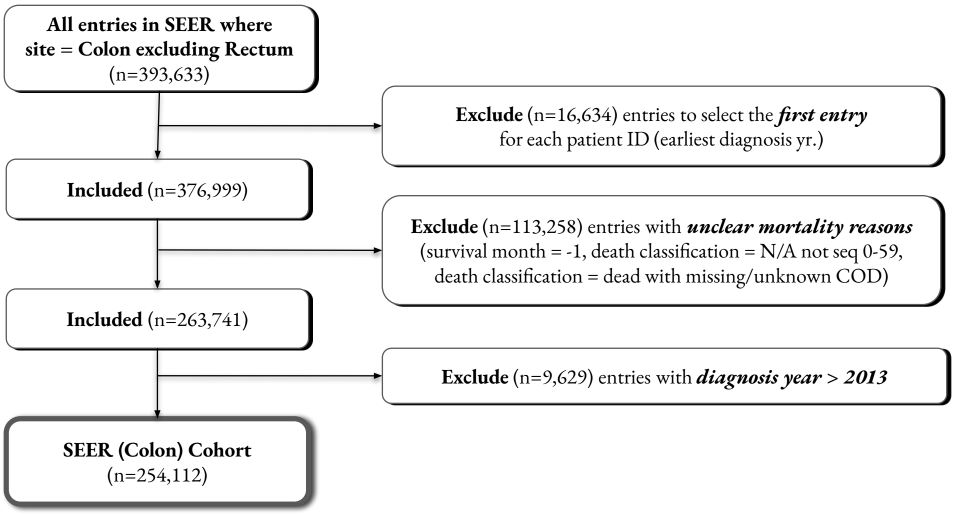

All entries in SEER where site = Colon excluding Rectum (n=393,633)
Exclude (n=16,634) entries to select the first entry for each patient ID (earliest diagnosis yr.)
Included (n=376,999)
Exclude (n=113,258) entries with unclear mortality reasons (survival month = -1, death classification = N/A not seq 0-59, death classification = dead with missing/unknown COD)
Included (n=263,741)
Exclude (n=9,629) entries with diagnosis year > 2013
SEER (Colon) Cohort (n=254,112)

Figure 8: Cohort selection diagram - SEER (Colon)

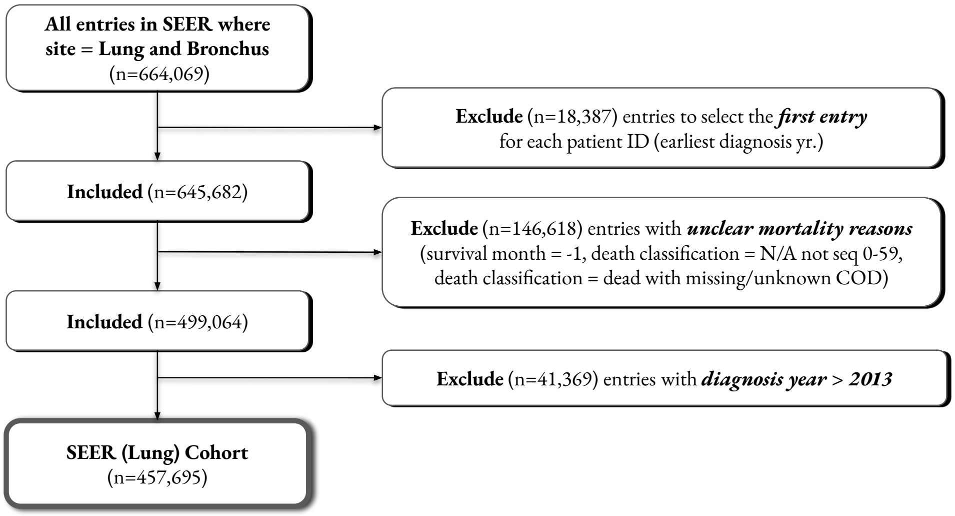

All entries in SEER where site = Lung and Bronchus (n=664,069)
Exclude (n=18,387) entries to select the ***first entry*** for each patient ID (earliest diagnosis yr.)
Included (n=645,682)
Exclude (n=146,618) entries with ***unclear mortality reasons*** (survival month = -1, death classification = N/A not seq 0-59, death classification = dead with missing/unknown COD)
Included (n=499,064)
Exclude (n=41,369) entries with ***diagnosis year*** > ***2013***
SEER (Lung) Cohort (n=457,695)

Figure 9: Cohort selection diagram - SEER (Lung)

Table 3: SEER (Breast) cohort characteristics, with count (%) or median (Q1 – Q3).

Characteristic		Missingness	Type
Sex			
Female	459,184 (99.4%)	–	categorical
Male	2,839 (0.6%)	–	categorical
Age recode with single ages and 85+	60 (50-71)	0.0%	continuous
Race/ethnicity			
White	387,247 (83.8%)	–	categorical
Black	40,217 (8.7%)	–	categorical
Other	34,559 (7.5%)	–	categorical
Laterality			
Right - origin of primary	224,777 (48.7%)	–	categorical
Left - origin of primary	233,549 (50.5%)	–	categorical
Other	3,697 (0.8%)	–	categorical
Regional nodes positive (1988+)	0 (0-3)	21.0%	continuous
T value - based on AJCC 3rd (1988-2003)	10 (10-20)	56.2%	categorical
Derived AJCC T, 7th ed (2010-2015)	13 (13-20)	85.3%	categorical
CS site-specific factor 3 (2004-2017 varying by schema)	0 (0-2)	64.8%	categorical
Regional nodes examined (1988+)	8 (2-15)	21.0%	continuous
Coding system-EOD (1973-2003)			
Four-digit EOD (1983-1987)	44,066 (9.5%)	–	categorical
Ten-digit EOD (1988-2003)	202,450 (43.8%)	–	categorical
Thirteen-digit (expanded) site specific EOD (1973-1982)	52,742 (11.4%)	–	categorical
Blank(s)	162,765 (35.2%)	–	categorical
CS version input original (2004-2015)	10,401 (10,300-20,302)	64.8%	categorical
CS version input current (2004-2015)	20,520 (20,510-20,540)	64.8%	categorical
EOD 10 - extent (1988-2003)	10 (10-13)	56.2%	categorical
Grade (thru 2017)			
Unknown	130,713 (28.3%)	–	categorical
Moderately differentiated; Grade II	135,970 (29.4%)	–	categorical
Poorly differentiated; Grade III	119,900 (26.0%)	–	categorical
Undifferentiated; anaplastic; Grade IV	8,081 (1.7%)	–	categorical
Well differentiated; Grade I	67,359 (14.6%)	–	categorical
SEER historic stage A (1973-2015)			
Regional	136,207 (29.5%)	–	categorical
Localized	286,927 (62.1%)	–	categorical
Unstaged	9,242 (2.0%)	–	categorical
Distant	29,647 (6.4%)	–	categorical
IHS Link			
Record sent for linkage, no IHS match	409,058 (88.5%)	–	categorical
Record sent for linkage, IHS match	1,505 (0.3%)	–	categorical
Blank(s)	51,460 (11.1%)	–	categorical
Histologic Type ICD-O-3	8,500 (8,500-8,500)	0.0%	categorical
EOD 10 - size (1988-2003)	18 (10-30)	56.2%	categorical
Type of Reporting Source			
Hospital inpatient/outpatient or clinic	450,801 (97.6%)	–	categorical
Other	11,222 (2.4%)	–	categorical
SEER cause-specific death classification			
Alive or dead of other cause	378,758 (82.0%)	–	categorical
Dead (attributable to this cancer dx)	83,265 (18.0%)	–	categorical
Survival months	135 (74-220)	0.0%	categorical
5-year survival			
1	378,758 (82.0%)	–	categorical
0	83,265 (18.0%)	–	categorical

Table 4: SEER (Colon) cohort characteristics, with count (%) or median (Q1–Q3).

Characteristic		Missingness	Type
Sex			
Female	133,661 (52.6%)	–	categorical
Male	120,451 (47.4%)	–	categorical
Age recode with single ages and 85+	70 (61-79)	0.0%	continuous
Race recode (White, Black, Other)			
White	212,265 (83.5%)	–	categorical
Black	24,041 (9.5%)	–	categorical
Other	17,806 (7.0%)	–	categorical
CS version input current (2004-2015)	20,510 (20,510-20,540)	72.8%	categorical
Derived AJCC T, 6th ed (2004-2015)	30 (20-40)	73.3%	categorical
Histology ICD-O-2	8,140 (8,140-8,210)	0.0%	categorical
IHS Link			
Record sent for linkage, no IHS match	208,802 (82.2%)	–	categorical
Record sent for linkage, IHS match	744 (0.3%)	–	categorical
Blank(s)	44,566 (17.5%)	–	categorical
Histology recode - broad groupings			
8140-8389: adenomas and adenocarcinomas	213,193 (83.9%)	–	categorical
8440-8499: cystic, mucinous and serous neoplasms	28,257 (11.1%)	–	categorical
8010-8049: epithelial neoplasms, NOS	8,797 (3.5%)	–	categorical
Other	3,865 (1.5%)	–	categorical
Regional nodes positive (1988+)	1 (0-10)	29.8%	continuous
CS mets at dx (2004-2015)	0 (0-22)	72.8%	continuous
Reason no cancer-directed surgery			
Surgery performed	223,929 (88.1%)	–	categorical
Not recommended	13,003 (5.1%)	–	categorical
Other	17,180 (6.8%)	–	categorical
Derived AJCC T, 6th ed (2004-2015)	30 (20-40)	73.3%	categorical
CS version input original (2004-2015)	10,401 (10,300-20,302)	72.8%	categorical
Primary Site	184 (182-187)	0.0%	categorical
Diagnostic Confirmation			
Positive histology	244,616 (96.3%)	–	categorical
Radiography without microscopic confirm	4,822 (1.9%)	–	categorical
Other	4,674 (1.8%)	–	categorical
EOD 10 - extent (1988-2003)	45 (40-85)	57.0%	categorical
Histologic Type ICD-O-3	8,140 (8,140-8,210)	0.0%	categorical
EOD 10 - size (1988-2003)	55 (35-999)	57.0%	categorical
CS lymph nodes (2004-2015)	0 (0-210)	72.8%	categorical
SEER cause-specific death classification			
Dead (attributable to this cancer dx)	119,047 (46.8%)	–	categorical
Alive or dead of other cause	135,065 (53.2%)	–	categorical
Survival months	68 (12-151)	0.0%	categorical
5-year survival			
1	135,065 (53.2%)	–	categorical
0	119,047 (46.8%)	–	categorical

Table 5: SEER (Lung) cohort characteristics, with count (%) or median (Q1 – Q3).

Characteristic		Missingness	Type
Sex			
Female	187,967 (41.1%)	–	categorical
Male	269,728 (58.9%)	–	categorical
Age recode with single ages and 85+	68 (60-76)	0.0%	continuous
Race recode (White, Black, Other)			
White	384,184 (83.9%)	–	categorical
Black	47,237 (10.3%)	–	categorical
Other	26,274 (5.7%)	–	categorical
Histologic Type ICD-O-3	8,070 (8,041-8,140)	0.0%	categorical
Laterality			
Left - origin of primary	178,661 (39.0%)	–	categorical
Right - origin of primary	245,321 (53.6%)	–	categorical
Paired site, but no information concerning laterality	23,196 (5.1%)	–	categorical
Other	10,517 (2.3%)	–	categorical
EOD 10 - nodes (1988-2003)	2 (1-9)	56.3%	categorical
EOD 4 - nodes (1983-1987)	3 (0-9)	88.4%	categorical
Type of Reporting Source			
Hospital inpatient/outpatient or clinic	445,606 (97.4%)	–	categorical
Other	12,089 (2.6%)	–	categorical
SEER historic stage A (1973-2015)			
Regional	79,409 (17.3%)	–	categorical
Distant	182,467 (39.9%)	–	categorical
Blank(s)	123,161 (26.9%)	–	categorical
Localized	50,375 (11.0%)	–	categorical
Unstaged	22,283 (4.9%)	–	categorical
CS version input current (2004-2015)	20,520 (20,510-20,540)	70.6%	categorical
CS mets at dx (2004-2015)	23 (0-40)	70.6%	continuous
CS version input original (2004-2015)	10,401 (10,300-20,302)	70.6%	categorical
CS tumor size (2004-2015)	50 (29-999)	70.6%	categorical
EOD 10 - size (1988-2003)	80 (35-999)	56.3%	categorical
CS lymph nodes (2004-2015)	200 (0-200)	70.6%	categorical
Histology recode - broad groupings			
8140-8389: adenomas and adenocarcinomas	147,127 (32.1%)	–	categorical
8010-8049: epithelial neoplasms, NOS	179,848 (39.3%)	–	categorical
8440-8499: cystic, mucinous and serous neoplasms	6,266 (1.4%)	–	categorical
Other	124,454 (27.2%)	–	categorical
EOD 10 - extent (1988-2003)	78 (40-85)	56.3%	categorical
SEER cause-specific death classification			
Alive or dead of other cause	49,997 (10.9%)	–	categorical
Dead (attributable to this cancer dx)	407,698 (89.1%)	–	categorical
Survival months	7 (2-19)	0.0%	categorical
5-year survival			
1	49,997 (10.9%)	–	categorical
0	407,698 (89.1%)	–	categorical

C.2. Features

SEER (Breast):

```
AJCC stage 3rd edition (1988-2003)
AYA site recode/WHO 2008
Age recode with single ages and 85+
Behavior code ICD-O-2
Behavior code ICD-O-3
Behavior recode for analysis
Breast - Adjusted AJCC 6th M (1988-2015)
Breast - Adjusted AJCC 6th N (1988-2015)
Breast - Adjusted AJCC 6th Stage (1988-2015)
Breast - Adjusted AJCC 6th T (1988-2015)
Breast Subtype (2010+)
CS Schema - AJCC 6th Edition
CS extension (2004-2015)
CS lymph nodes (2004-2015)
CS mets at dx (2004-2015)
CS site-specific factor 1 (2004-2017 varying by schema)
CS site-specific factor 15 (2004-2017 varying by schema)
CS site-specific factor 2 (2004-2017 varying by schema)
CS site-specific factor 25 (2004-2017 varying by schema)
CS site-specific factor 3 (2004-2017 varying by schema)
CS site-specific factor 4 (2004-2017 varying by schema)
CS site-specific factor 5 (2004-2017 varying by schema)
CS site-specific factor 6 (2004-2017 varying by schema)
CS site-specific factor 7 (2004-2017 varying by schema)
CS tumor size (2004-2015)
CS version derived (2004-2015)
CS version input current (2004-2015)
CS version input original (2004-2015)
Coding system-EOD (1973-2003)
Derived AJCC M, 6th ed (2004-2015)
Derived AJCC M, 7th ed (2010-2015)
Derived AJCC N, 6th ed (2004-2015)
Derived AJCC N, 7th ed (2010-2015)
Derived AJCC Stage Group, 6th ed (2004-2015)
Derived AJCC Stage Group, 7th ed (2010-2015)
Derived AJCC T, 6th ed (2004-2015)
Derived AJCC T, 7th ed (2010-2015)
Derived HER2 Recode (2010+)
EOD 10 - extent (1988-2003)
EOD 10 - nodes (1988-2003)
EOD 10 - size (1988-2003)
ER Status Recode Breast Cancer (1990+)
First malignant primary indicator
Grade (thru 2017)
Histologic Type ICD-O-3
Histology recode - Brain groupings
Histology recode - broad groupings
ICCC site rec extended ICD-O-3/WHO 2008
IHS Link
Laterality
Lymphoma subtype recode/WHO 2008 (thru 2017)
M value - based on AJCC 3rd (1988-2003)
N value - based on AJCC 3rd (1988-2003)
Origin recode NHIA (Hispanic, Non-Hisp)
PR Status Recode Breast Cancer (1990+)
Primary Site
Primary by international rules
Race recode (W, B, AI, API)
Race recode (White, Black, Other)
Race/ethnicity
Regional nodes examined (1988+)
Regional nodes positive (1988+)
SEER historic stage A (1973-2015)
SEER modified AJCC stage 3rd (1988-2003)
Sex
Site recode ICD-O-3/WHO 2008
T value - based on AJCC 3rd (1988-2003)
Tumor marker 1 (1990-2003)
Tumor marker 2 (1990-2003)
Tumor marker 3 (1998-2003)
Type of Reporting Source
```

SEER (Colon):

```
Age recode with <1 year olds
Age recode with single ages and 85+
Behavior code ICD-O-2
Behavior code ICD-O-3
CS extension (2004-2015)
CS lymph nodes (2004-2015)
CS mets at dx (2004-2015)
CS site-specific factor 1 (2004-2017 varying by schema)
CS tumor size (2004-2015)
CS version input current (2004-2015)
CS version input original (2004-2015)
Derived AJCC M, 6th ed (2004-2015)
Derived AJCC M, 7th ed (2010-2015)
Derived AJCC N, 6th ed (2004-2015)
Derived AJCC N, 7th ed (2010-2015)
Derived AJCC Stage Group, 6th ed (2004-2015)
Derived AJCC Stage Group, 7th ed (2010-2015)
Derived AJCC T, 6th ed (2004-2015)
Derived AJCC T, 7th ed (2010-2015)
Diagnostic Confirmation
EOD 10 - extent (1988-2003)
EOD 10 - nodes (1988-2003)
EOD 10 - size (1988-2003)
Histologic Type ICD-O-3
Histology ICD-O-2
Histology recode - broad groupings
IHS Link
Origin recode NHIA (Hispanic, Non-Hisp)
Primary Site
Primary by international rules
RX Summ--Surg Prim Site (1998+)
Race recode (White, Black, Other)
Reason no cancer-directed surgery
Regional nodes positive (1988+)
SEER modified AJCC stage 3rd (1988-2003)
Sex
```

SEER (Lung):

```
AYA site recode/WHO 2008
Age recode with <1 year olds
Age recode with single ages and 85+
Behavior code ICD-O-2
Behavior code ICD-O-3
CS extension (2004-2015)
CS lymph nodes (2004-2015)
CS mets at dx (2004-2015)
CS site-specific factor 1 (2004-2017 varying by schema)
CS tumor size (2004-2015)
CS version input current (2004-2015)
CS version input original (2004-2015)
Derived AJCC M, 6th ed (2004-2015)
Derived AJCC M, 7th ed (2010-2015)
Derived AJCC N, 6th ed (2004-2015)
Derived AJCC N, 7th ed (2010-2015)
Derived AJCC Stage Group, 6th ed (2004-2015)
Derived AJCC T, 6th ed (2004-2015)
Derived AJCC T, 7th ed (2010-2015)
EOD 10 - extent (1988-2003)
EOD 10 - nodes (1988-2003)
EOD 10 - size (1988-2003)
EOD 4 - nodes (1983-1987)
First malignant primary indicator
Grade (thru 2017)
Histologic Type ICD-O-3
Histology recode - broad groupings
ICCC site recode 3rd edition/IARC 2017
ICCC site recode extended 3rd edition/IARC 2017
IHS Link
Laterality
Origin recode NHIA (Hispanic, Non-Hisp)
Primary by international rules
Race recode (White, Black, Other)
SEER historic stage A (1973-2015)
Sex
Type of Reporting Source
```

C.3. Missingness heatmaps

This section plots missingness heatmaps of categorical and numerical features in each SEER dataset over time. Darker color means larger proportion of missing data.

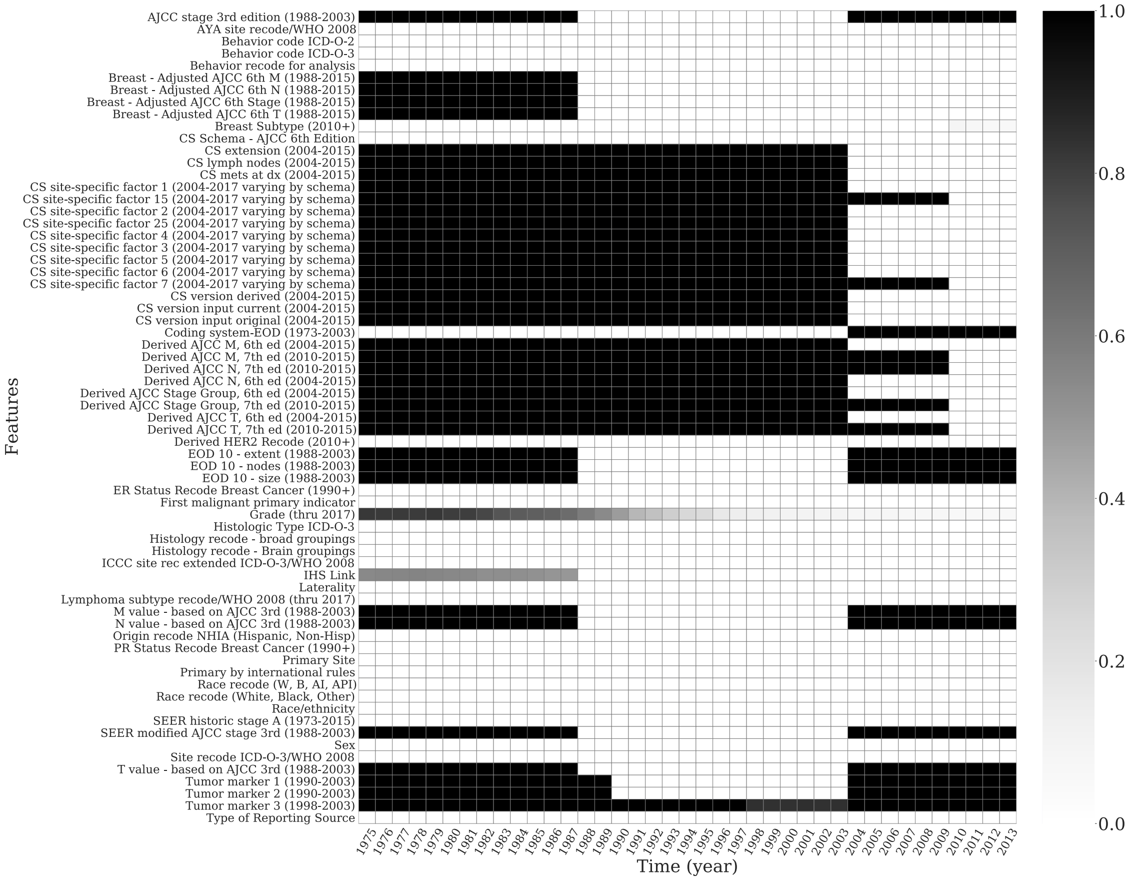


Figure 10: Missingness of categorical features in SEER (Breast).

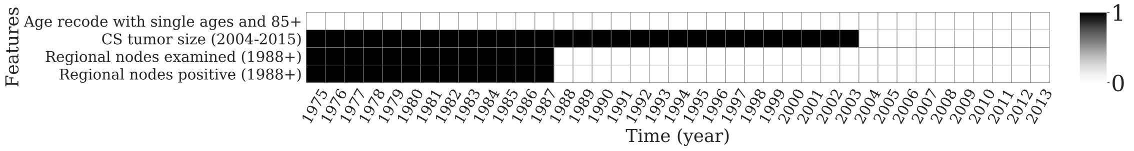


Figure 11: Missingness of numerical features in SEER (Breast).

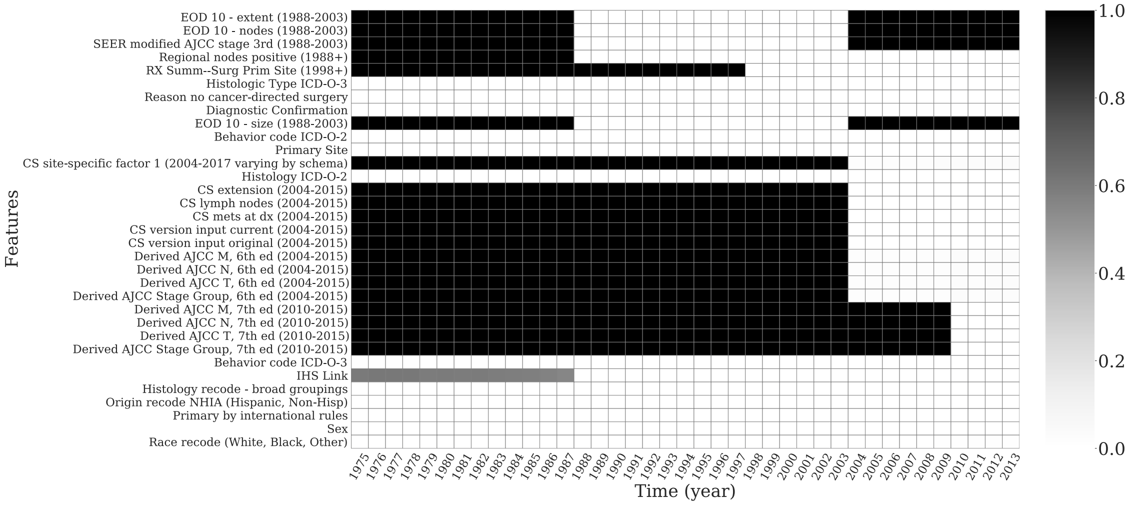


Figure 12: Missingness of categorical features in SEER (Colon).

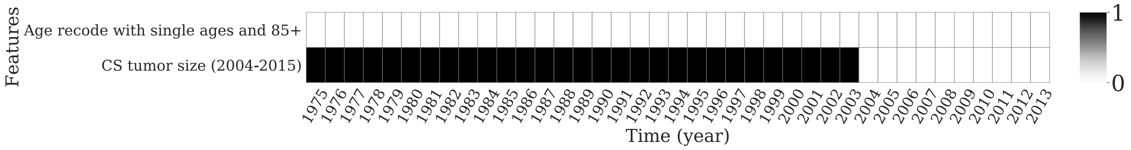


Figure 13: Missingness of numerical features in SEER (Colon).

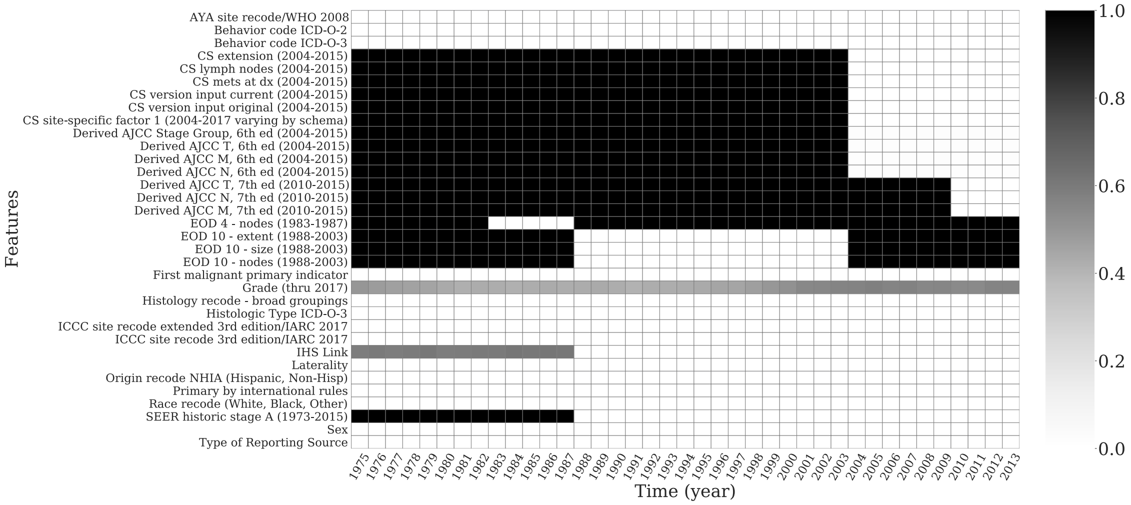


Figure 14: Missingness of categorical features in SEER (Lung).

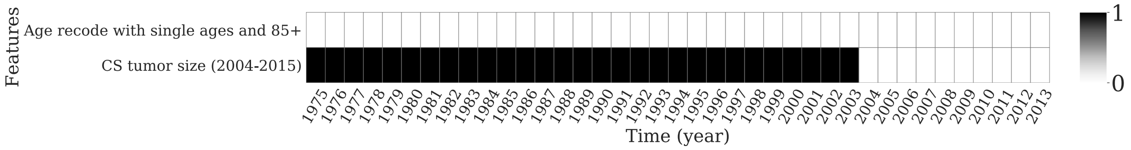


Figure 15: Missingness of numerical features in SEER (Lung).

Appendix D. Additional CDC COVID-19 Data Details

The COVID-19 Case Surveillance Detailed Data (Centers for Disease Control and Prevention, 2020) is a national, publicly available dataset provided by the CDC. It contains 33 elements, with patient-level data including symptoms, demographics, and state of residence. The performance over time is evaluated on a *monthly* basis. We use the version the released on June 6th, 2022. Disclaimer: "The CDC does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented."

- Data access: To access the data, users must complete a registration information and data use restrictions agreement (RIDURA).
- Cohort selection: The cohort consists of all patients who were lab-confirmed positive for COVID-19, had a non-null positive specimen date, and were hospitalized (`hosp_yn = Yes`). Cohort selection diagrams are given in Figures 16
- Cohort characteristics: Cohort characteristics are given in Table 6.
- Outcome definition: mortality, defined by `death_yn = Yes`
- Features: We list the features used in the CDC COVID-19 datasets in Section D.2. We convert all categorical variables into dummy features, and apply standard scaling to numerical variables (subtract mean and divide by standard deviation).
- Missingness heat map: is given in Figure 17.
- Additionally, we provide stacked area plots showing how the distribution of ages (Figure 18(a) and states 18(b) shifts over time.

D.1. Cohort Selection and Cohort Characteristics

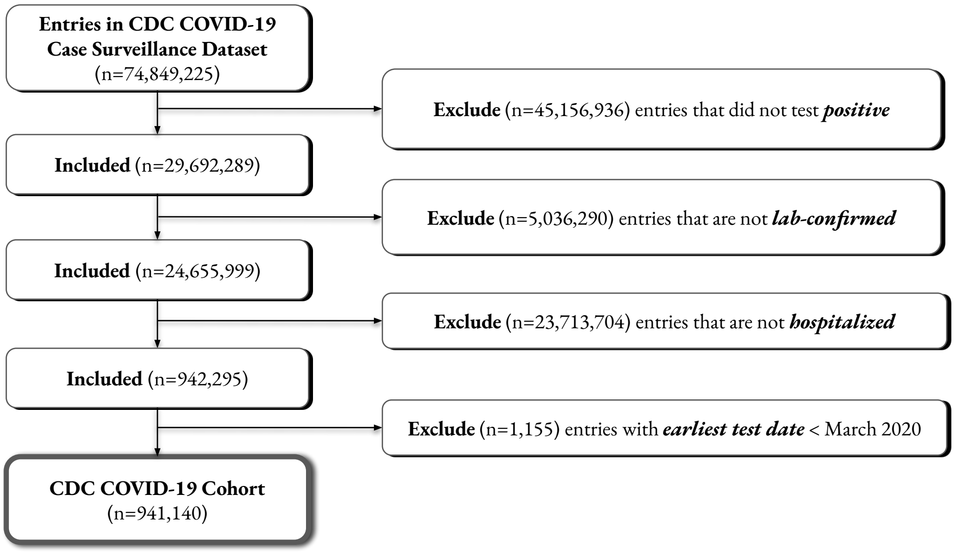

Entries in CDC COVID-19 Case Surveillance Dataset (n=74,849,225)
Exclude (n=45,156,936) entries that did not test positive
Included (n=29,692,289)
Exclude (n=5,036,290) entries that are not lab-confirmed
Included (n=24,655,999)
Exclude (n=23,713,704) entries that are not hospitalized
Included (n=942,295)
Exclude (n=1,155) entries with earliest test date < March 2020
CDC COVID-19 Cohort (n=941,140)

Figure 16: Cohort selection diagram - CDC COVID-19

Table 6: CDC COVID-19 cohort characteristics, with count (%) or median (Q1–Q3).

Characteristic		Missingness	Type
Sex			
Female	455,376 (48.4%)	–	categorical
Male	475,223 (50.5%)	–	categorical
Unknown/Missing	10,541 (1.1%)	–	categorical
Age Group			
0 - 9	16,373 (1.7%)	–	categorical
10 - 19	17,252 (1.8%)	–	categorical
20 - 29	48,505 (5.2%)	–	categorical
30 - 39	71,776 (7.6%)	–	categorical
40 - 49	88,531 (9.4%)	–	categorical
50 - 59	141,805 (15.1%)	–	categorical
60 - 69	189,354 (20.1%)	–	categorical
70 - 79	189,018 (20.1%)	–	categorical
80+	177,765 (18.9%)	–	categorical
Missing	761 (0.1%)	–	categorical
Race			
White	544,199 (57.8%)	–	categorical
Black	173,847 (18.5%)	–	categorical
Other	205,547 (21.8%)	–	categorical
State of Residence			
NY	189,684 (20.2%)	–	categorical
OH	70,097 (7.4%)	–	categorical
FL	35,679 (3.8%)	–	categorical
WA	58,854 (6.3%)	–	categorical
MA	31,441 (3.3%)	–	categorical
Other	555,353 (59.0%)	–	categorical
Mechanical Ventilation			
Yes	38,009 (4.0%)	–	categorical
No	138,331 (14.7%)	–	categorical
Unknown/Missing	764,800 (81.2%)	–	categorical
Mortality			
1	190,786 (20.3%)	–	categorical
0	750,354 (79.7%)	–	categorical

D.2. Features

`abdom_yn, abxchest_yn, acuterespdistress_yn, age_group, chills_yn, cough_yn, diarrhea_yn, ethnicity, fever_yn, hc_work_yn, headache_yn, hosp_yn, icu_yn, mechvent_yn, medcond_yn, month, myalgia_yn, nauseavomit_yn, pna_yn, race, relative_month, res_county, res_state, runnose_yn, sex, sfever_yn, sob_yn, sthroat_yn,`

D.3. Missingness heatmaps

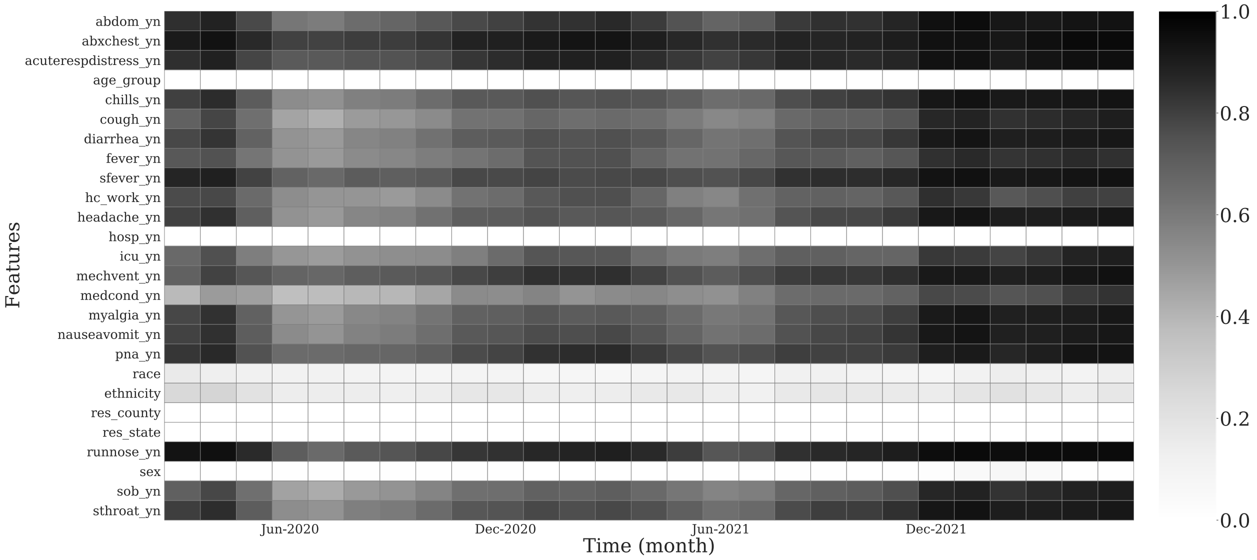
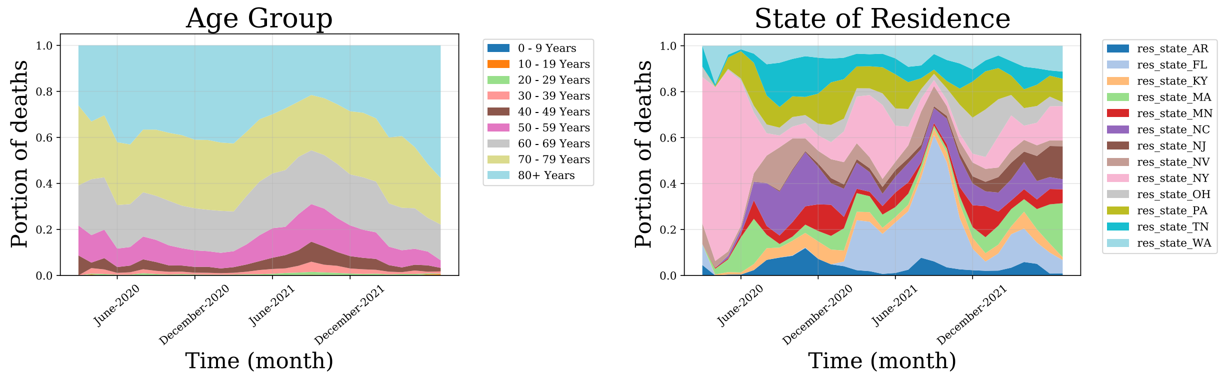


Figure 17: Missingness over time for features in CDC COVID-19 dataset after cohort selection. The darker the color, the larger the proportion of missing data.

D.4. Additional Figures



(*a*) By age group

(*b*) By state of residence

Figure 18: Proportion of deaths over time for each age group and state of residence.

Appendix E. Additional SWPA COVID-19 Data Details

The Southwestern Pennsylvania (SWPA) COVID-19 dataset consists of EHR data from patients tested for COVID-19. It was collected by a major healthcare provider in SWPA, and includes patient demographics, labs, problem histories, medications, inpatient vs. outpatient status, and other information collected in the patient encounter. The performance over time is evaluated on a *monthly* basis.

- Data access: This is a private dataset.
- Cohort selection: The cohort consists of COVID-19 patients who tested positive for COVID-19 and were not already in the ICU or mechanically ventilated. We filter for the first positive test, and define features and outcomes relative to that time. Cohort selection diagrams are given in Figures 19. If there are multiple samples per patient, we filter to the first entry per patient, which corresponds to when a patient first enters the dataset. This corresponds to a particular interpretation of the prediction: when a patient is first tests positive, given what we know about that patient, what is their estimated risk of 90-day mortality?
- Cohort characteristics: Cohort characteristics are given in Table 7.
- Outcome definition: 90-day mortality by comparing the death date and test date
- Features: We list the features used in the SWPA COVID-19 datasets in Section E.2. We convert all categorical variables into dummy features, and apply standard scaling to numerical variables (subtract mean and divide by standard deviation). To create a fixed length feature vector, where applicable we take the most recent value of each feature (e.g. most recent lab values).
- Missingness heat maps: are given in Figures 20, 21, 22, and 23,

E.1. Cohort Selection and Cohort Characteristics

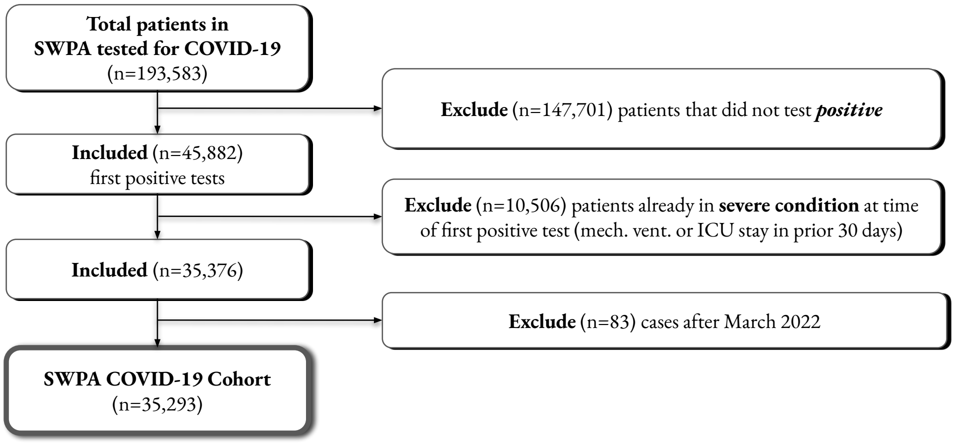

Total patients in SWPA tested for COVID-19 (n=193,583)
Exclude (n=147,701) patients that did not test ***positive***
Included (n=45,882) first positive tests
Exclude (n=10,506) patients already in **severe condition** at time of first positive test (mech. vent. or ICU stay in prior 30 days)
Included (n=35,376)
Exclude (n=83) cases after March 2022
SWPA COVID-19 Cohort (n=35,293)

Figure 19: Cohort selection diagram - SWPA COVID-19

Table 7: SWPA COVID-19 cohort characteristics, with count (%) or median (Q1–Q3).

Characteristic		Missingness	Type
Gender			
Female	20,283 (57.5%)	–	categorical
Male	15,003 (42.5%)	–	categorical
Unknown	7 (0.0%)	–	categorical
Age			
Under 20	3,210 (9.1%)	–	categorical
20 – 30	4,349 (12.3%)	–	categorical
30 – 40	4,667 (13.2%)	–	categorical
40 – 50	4,653 (13.2%)	–	categorical
50 – 60	6,111 (17.3%)	–	categorical
60 – 70	5,700 (16.2%)	–	categorical
70+	6,603 (18.7%)	–	categorical
Location of test			
Inpatient	14,911 (42.2%)	–	categorical
Outpatient	17,661 (50.0%)	–	categorical
Unknown	2,721 (7.7%)	–	categorical
90-day mortality			
True	1,516 (4.3%)	–	categorical
False	33,777 (95.7%)	–	categorical

E.2. Features

```
Asthma
CAD
CHF
CKD
COPD
CRP
CVtest_ICD_Acute pharyngitis, unspecified
CVtest_ICD_Acute upper respiratory infection, unspecified
CVtest_ICD_Anosmia
CVtest_ICD_Contact with and (suspected) exposure to other viral
    communicable diseases
CVtest_ICD_Encounter for general adult medical
    examination without
    abnormal findings
CVtest_ICD_Encounter for screening for other viral diseases
CVtest_ICD_Encounter for screening for respiratory disorder NEC
CVtest_ICD_Nasal congestion
CVtest_ICD_Other general symptoms and signs
CVtest_ICD_Other specified symptoms and signs involving the
    circulatory and respiratory systems
CVtest_ICD_Pain, unspecified
CVtest_ICD_Parageusia
CVtest_ICD_R05.9
CVtest_ICD_R51.9
CVtest_ICD_U07.1
CVtest_ICD_Viral infection, unspecified
CVtest_ICD_Z20.822
ESLD
Hypertension
IP_ICD_z20.828
Immunocompromised
Interstitial Lung disease
OP_ICD_Abdominal Pain
OP_ICD_Chest Pain
OP_ICD_Chills
OP_ICD_Coronavirus Concerns
OP_ICD_Covid Infection
OP_ICD_Exposure To Covid-19
OP_ICD_Generalized Body Aches
OP_ICD_Headache
OP_ICD_Labs Only
OP_ICD_Medication Refill
OP_ICD_Nasal Congestion
OP_ICD_Nausea
OP_ICD_Other
OP_ICD_Results
OP_ICD_Shortness of Breath
OP_ICD_Sore Throat
OP_ICD_URI
age_bin_(20, 30]
age_bin_(30, 40]
age_bin_(40, 50]
age_bin_(50, 60]
age_bin_(60, 70]
age_bin_(70, 200]
bmi
cancer
cough
covid_vaccination_given
diabetes
fatigue
fever
gender
hyperglycemia
lab_ANION GAP
lab_ATRIAL RATE
lab_BASOPHILS ABSOLUTE COUNT
lab_BASOPHILS RELATIVE PERCENT
lab_BLOOD UREA NITROGEN
lab_CALCIUM
lab_CALCUALTED T AXIS
lab_CALCULATED R AXIS
lab_CHLORIDE
lab_CO2
lab_CREATININE
lab_EOSINOPHILS ABSOLUTE COUNT
lab_EOSINOPHILS RELATIVE PERCENT
lab_GFR MDRD AF AMER
lab_GFR MDRD NON AF AMER
lab_GLUCOSE
lab_IMMATURE GRANULOCYTES RELATIVE PERCENT
lab_LYMPHOCYTES ABSOLUTE COUNT
lab_LYMPHOCYTES RELATIVE PERCENT
lab_MEAN CORPUSCULAR HEMOGLOBIN
lab_MEAN CORPUSCULAR HEMOGLOBIN CONC
lab_MEAN PLATELET VOLUME
lab_MONOCYTES ABSOLUTE COUNT
lab_MONOCYTES RELATIVE PERCENT
lab_NEUTROPHILS RELATIVE PERCENT
lab_NUCLEATED RED BLOOD CELLS
lab_POTASSIUM
lab_PROTEIN TOTAL
lab_Q-T INTERVAL
lab_QRS DURATION
lab_QTC CALCULATION
lab_RED CELL DISTRIBUTION WIDTH
lab_SODIUM
lab_VENTRICULAR RATE
lab_merged_CRP
lab_merged_albumin
lab_merged_alkalinePhosphatase
lab_merged_alt
lab_merged_ast
lab_merged_bnp
lab_merged_ddimer
lab_merged_directBilirubin
lab_merged_ggt
lab_merged_hct
lab_merged_hgb
lab_merged_indirectBilirubin
lab_merged_lactate
lab_merged_ldh
lab_merged_mcv
lab_merged_neutrophil
lab_merged_platelets
lab_merged_pt
lab_merged_rbc
lab_merged_sao2
lab_merged_totalBilirubin
lab_merged_totalProtein
lab_merged_troponin
lab_merged_wbc
labs_ICD_Acute pharyngitis, unspecified
labs_ICD_Acute upper respiratory infection, unspecified
labs_ICD_Chest pain, unspecified
labs_ICD_Contact with and (suspected) exposure to other
    viral communicable diseases
labs_ICD_Dyspnea, unspecified
labs_ICD_Encounter for other preprocedural examination
labs_ICD_Essential (primary) hypertension
labs_ICD_Fever, unspecified
labs_ICD_Heart failure, unspecified
labs_ICD_Other general symptoms and signs
labs_ICD_Other pulmonary embolism without acute cor pulmonale
labs_ICD_Other specified abnormalities of plasma proteins
labs_ICD_R05.9
labs_ICD_Shortness of breath
labs_ICD_Syncope and collapse
labs_ICD_U07.1
labs_ICD_Unspecified atrial fibrillation
labs_ICD_Viral infection, unspecified
labs_ICD_Z20.822
liver disease
location_covidtest_ordered_Inpatient
location_covidtest_ordered_Outpatient
lung disease
med_dx_Acquired hypothyroidism
med_dx_Anxiety
med_dx_COVID-19
med_dx_Encounter for antineoplastic chemotherapy
med_dx_Encounter for antineoplastic chemotherapy and immunotherapy
med_dx_Encounter for antineoplastic immunotherapy
med_dx_Encounter for immunization
med_dx_Gastroesophageal reflux disease without esophagitis
med_dx_Gastroesophageal reflux disease, esophagitis presence
    not specified
med_dx_Generalized anxiety disorder
med_dx_Hyperlipidemia, unspecified hyperlipidemia type
med_dx_Hypomagnesemia
med_dx_Hypothyroidism, unspecified type
med_dx_Iron deficiency anemia, unspecified iron deficiency anemia type
med_dx_Mixed hyperlipidemia
med_dx_Primary osteoarthritis of right knee
medication_ACETAMINOPHEN 325 MG TABLET
medication_ALBUTEROL SULFATE 2.5 MG/3 ML (0.083 %) SOLUTION
    FOR NEBULIZATION
medication_ALBUTEROL SULFATE HFA 90 MCG/ACTUATION AEROSOL INHALER
medication_ASPIRIN 81 MG TABLET,DELAYED RELEASE
medication_DEXAMETHASONE SODIUM PHOSPHATE 4 MG/ML INJECTION SOLUTION
medication_DIPHENHYDRAMINE 50 MG/ML INJECTION (WRAPPER)
medication_EPINEPHRINE 0.3 MG/0.3 ML INJECTION, AUTO-INJECTOR
medication_FENTANYL (PF) 50 MCG/ML INJECTION SOLUTION
medication_HYDROCODONE 5 MG-ACETAMINOPHEN 325 MG TABLET
medication_HYDROCORTISONE SOD SUCCINATE (PF) 100 MG/2 ML SOLUTION
    FOR INJECTION
medication_IOPAMIDOL 76 % INTRAVENOUS SOLUTION
medication_LACTATED RINGERS INTRAVENOUS SOLUTION
medication_MIDAZOLAM 1 MG/ML INJECTION SOLUTION
medication_NALOXONE 0.4 MG/ML INJECTION SOLUTION
medication_ONDANSETRON HCL (PF) 4 MG/2 ML INJECTION SOLUTION
medication_OXYCODONE 5 MG TABLET
medication_PANTOPRAZOLE 40 MG TABLET,DELAYED RELEASE
medication_PROPOFOL 10 MG/ML INTRAVENOUS BOLUS (20 ML)
medication_SODIUM CHLORIDE 0.9 % INTRAVENOUS SOLUTION
medication_SODIUM CHLORIDE 0.9 % IV BOLUS
myalgia
obesity
past7Dprobhx_ICD_Acute kidney failure, unspecified
past7Dprobhx_ICD_Anemia, unspecified
past7Dprobhx_ICD_Anxiety disorder, unspecified
past7Dprobhx_ICD_Chest pain, unspecified
past7Dprobhx_ICD_Dizziness and giddiness
past7Dprobhx_ICD_Encounter for general adult medical examination
    without abnormal findings
past7Dprobhx_ICD_Encounter for immunization
past7Dprobhx_ICD_Encounter for screening for malignant
    neoplasm of colon
past7Dprobhx_ICD_F32.A
past7Dprobhx_ICD_Gastro-esophageal reflux disease
    without esophagitis
```

```
past7Dprobhx_ICD_Hyperlipidemia, unspecified
past7Dprobhx_ICD_Hypokalemia
past7Dprobhx_ICD_Hypothyroidism, unspecified
past7Dprobhx_ICD_Mixed hyperlipidemia
past7Dprobhx_ICD_Obstructive sleep apnea (adult) (pediatric)
past7Dprobhx_ICD_Syncope and collapse
past7Dprobhx_ICD_Type 2 diabetes mellitus without complications
past7Dprobhx_ICD_Unspecified atrial fibrillation
probhx_ICD_Acute kidney failure, unspecified
probhx_ICD_Anemia, unspecified
probhx_ICD_Anxiety disorder, unspecified
probhx_ICD_Chest pain, unspecified
probhx_ICD_Dizziness and giddiness
probhx_ICD_Encounter for general adult medical examination without
    abnormal findings
probhx_ICD_Encounter for immunization
probhx_ICD_Encounter for screening for malignant neoplasm of colon
probhx_ICD_F32.A
probhx_ICD_Gastro-esophageal reflux disease without esophagitis
probhx_ICD_Hyperlipidemia, unspecified
probhx_ICD_Hypokalemia
probhx_ICD_Hypothyroidism, unspecified
probhx_ICD_Mixed hyperlipidemia
probhx_ICD_Obstructive sleep apnea (adult) (pediatric)
probhx_ICD_Syncope and collapse
probhx_ICD_Type 2 diabetes mellitus without complications
probhx_ICD_Unspecified atrial fibrillation
transplant
troponin
vaccine_COVID-19 RS-AD26 (PF) Vaccine (Janssen)
vaccine_COVID-19 Vaccine, Unspecified
vaccine_COVID-19 mRNA (PF) Vaccine (Moderna)
vaccine_COVID-19 mRNA (PF) Vaccine (Pfizer)
vaccine_Flu Whole
vaccine_INFLUENZA, CCIV4
vaccine_Influenza
vaccine_Influenza High PF
vaccine_Influenza ID PF
vaccine_Influenza PF
vaccine_Influenza Vaccine, Quadrivalent, Adjuvanted
vaccine_Influenza, High-dose, Quadrivalent
vaccine_Influenza, Quadrivalent
vaccine_Influenza, Recombinant (RIV4)
vaccine_Influenza, Recombinant (Riv3)
vaccine_Influenza, Trivalent, Adjuvanted
vaccine_LAIV3
vaccine_Pneumococcal
vaccine_Pneumococcal Conjugate 13-valent
vaccine_Pneumococcal Polysaccharide
vaccine_TIVA
```

E.3. Missingness heatmaps

This section plots missingness heatmaps of categorical and numerical features over time. Darker color means larger proportion of missing data.

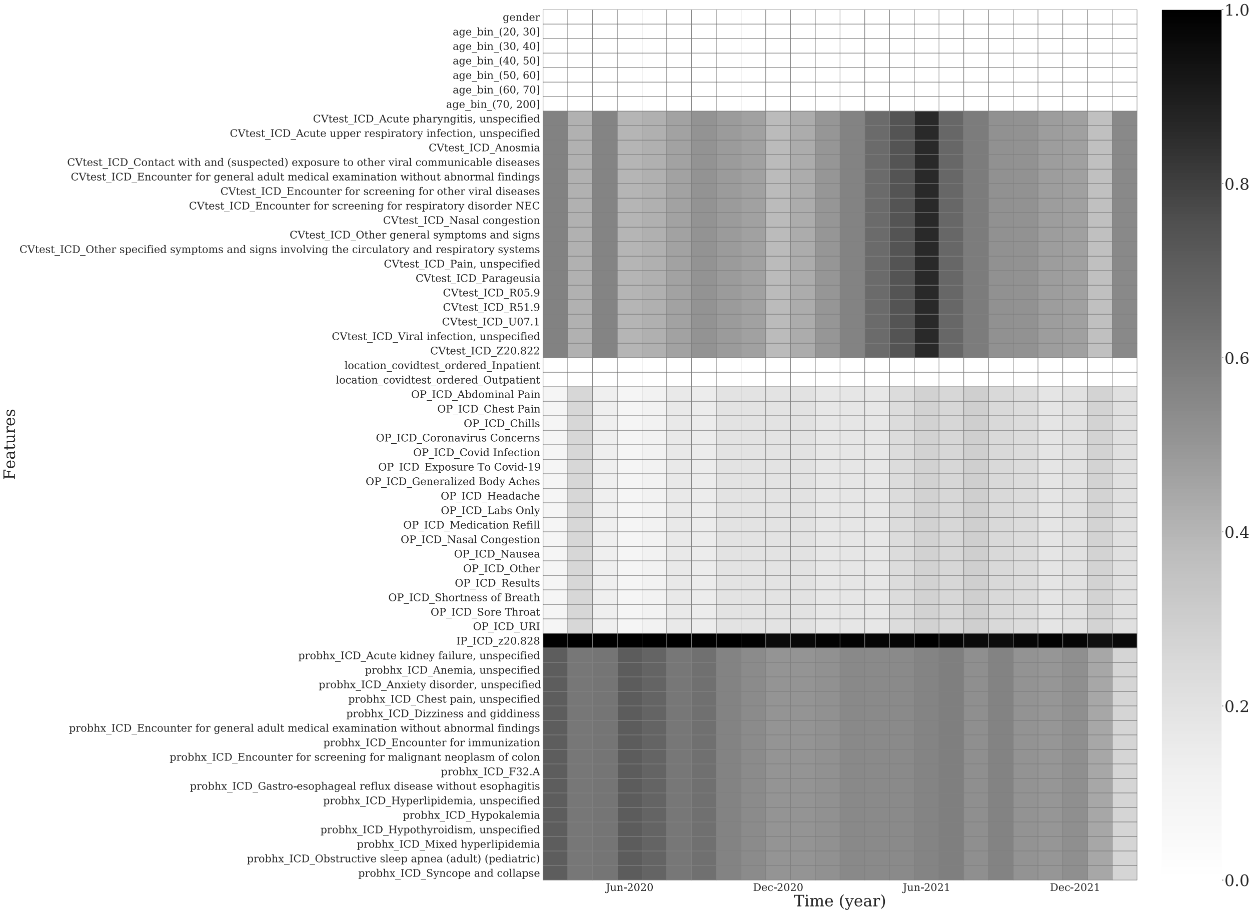


Figure 20: Missingness of categorical features in SWPA COVID-19 dataset (part 1).

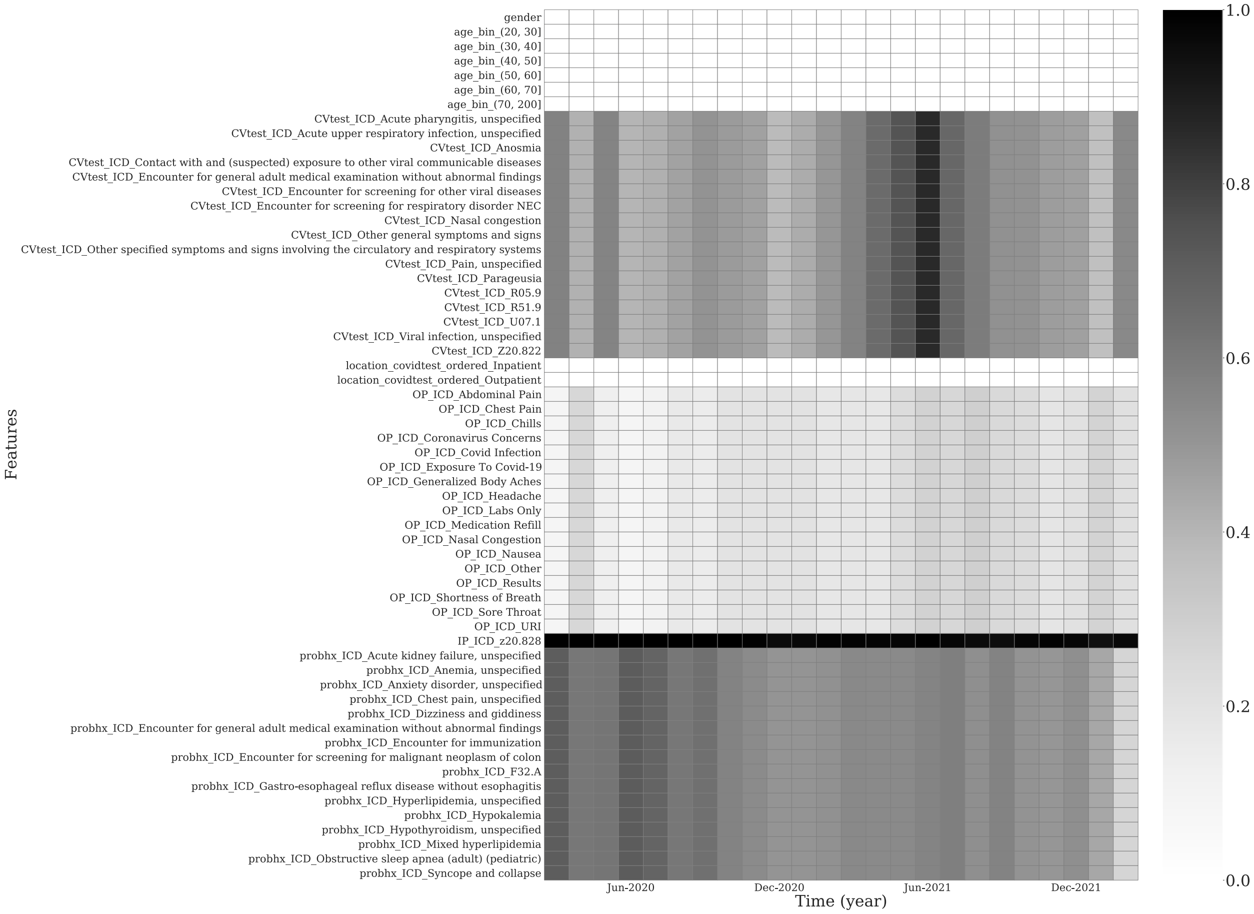


Figure 21: Missingness of categorical features in SWPA COVID-19 dataset (part 2).

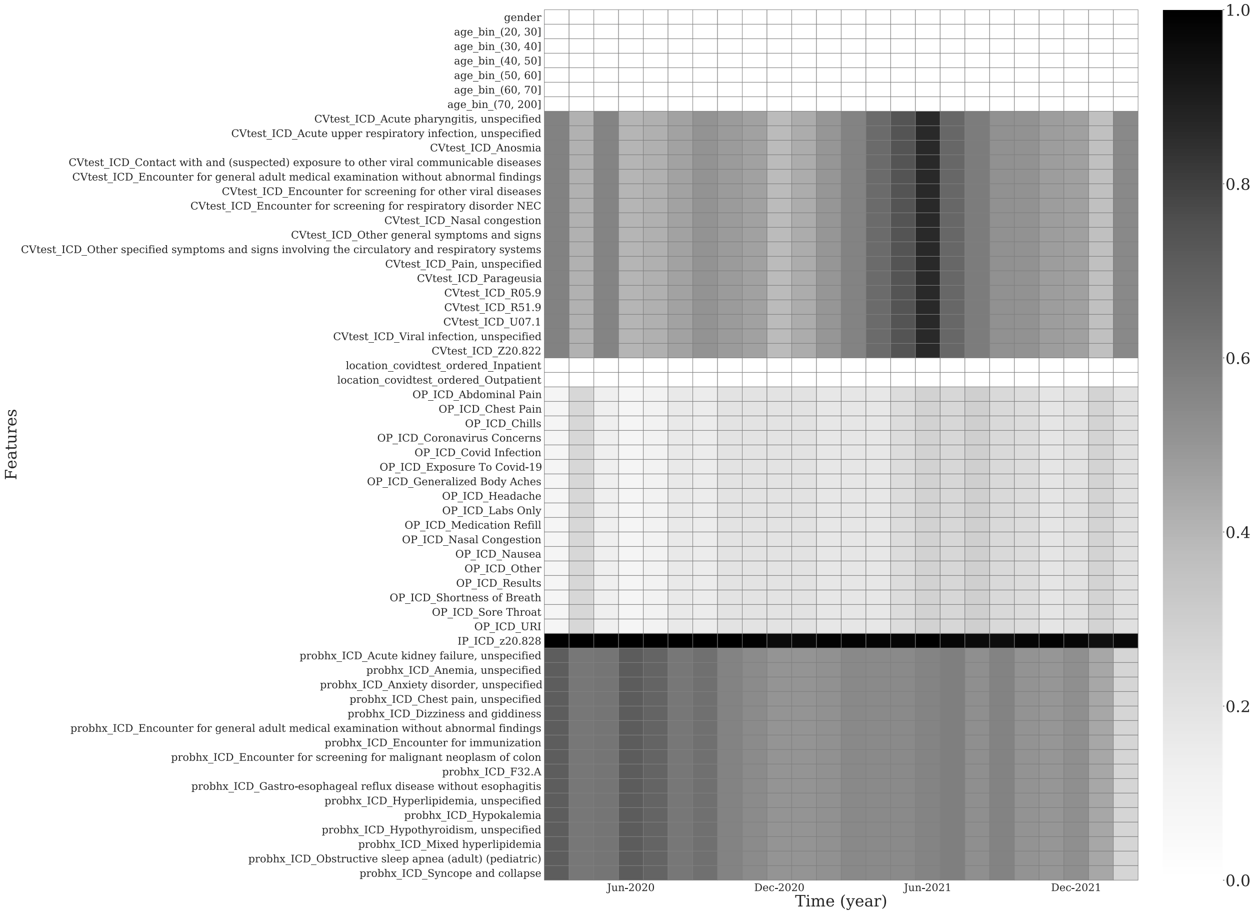


Figure 22: Missingness of categorical features in SWPA COVID-19 dataset (part 3).

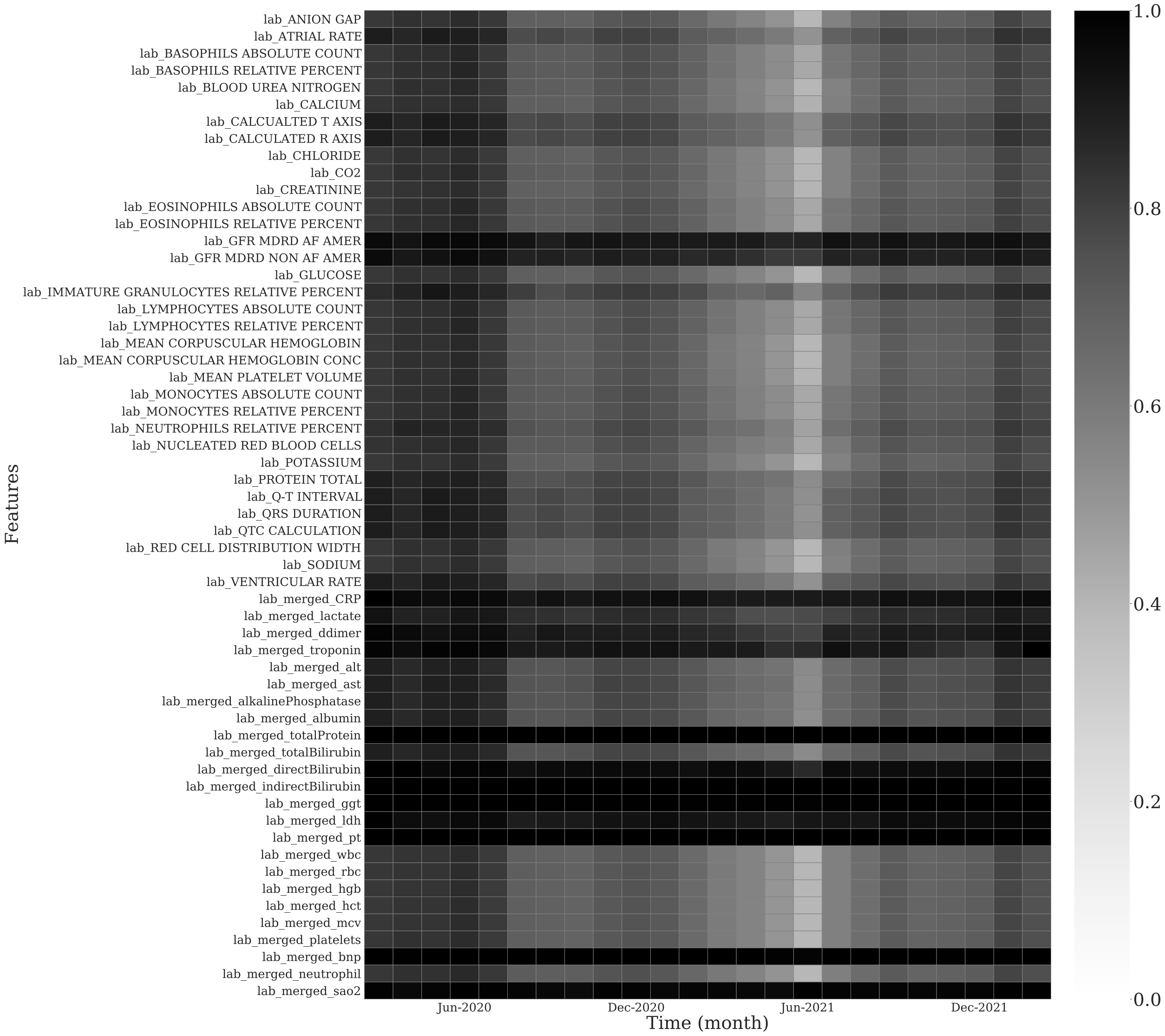


Figure 23: Missingness of numerical features in SWPA COVID-19.

Appendix F. Additional MIMIC-IV Data Details

The Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2021) database contains EHR data from patients admitted to critical care units from 2008–2019. MIMIC-IV is an update to MIMIC-III, adding time annotations placing each sample into a three-year time range, and removing elements from the old CareVue EHR system (before 2008). Each patient has an `anchor_year_group`, `anchor_year` and `intime`. For each patient, we first calculated an offset as the difference between `intime` and `anchor_year`. Then, we approximated the admit time as the midpoint of `anchor_year_group` after applying the computed offset.

The performance over time is evaluated on a *yearly* basis. Our study uses MIMIC-IV-1.0.

- Data access: Users must create a Physionet account, become credentialed, and sign a data use agreement (DUA).
- Cohort selection: We select all patients in the `icustays` table, filtering for their first encounter (minimum `intime`), and defining a feature vector only using information available by the first 24 hrs of their first encounter. (Selection diagram in Figure 24). If there are multiple samples per patient, we filter to the first entry per patient, which corresponds to when a patient first enters the dataset. This corresponds to a particular interpretation of the prediction: when a patient first visits the ICU, given what we know about that patient, what is their estimated risk of in-ICU mortality?
- Outcome definition: The outcome of interest is in-ICU mortality, defined by comparing the `outtime` of the patient's ICU visit with the patient's `dod` (date of death, in the `patients` table). As noted in the documentation, out-of-hospital mortality is not recorded.
- Cohort characteristics: Cohort characteristics are given in Table 8.
- Features: We list the features used in the MIMIC-IV datasets in Section F.2. We convert all categorical variables into dummy features, and apply standard scaling to numerical variables (subtract mean and divide by standard deviation). To create a fixed length feature vector, we take the most recent value of any patient history data available (e.g. most recent lab values).
- Missingness heat maps: are given in Figures 25, 26, 27, 28.

F.1. Cohort Selection and Cohort Characteristics

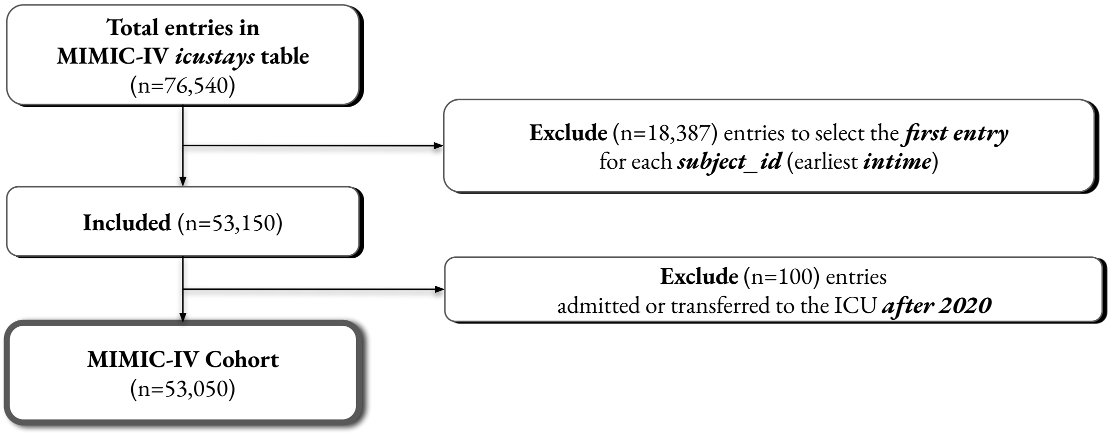

Total entries in
MIMIC-IV *icustays* table
(n=76,540)
Exclude (n=18,387) entries to select the ***first entry***
for each ***subject_id*** (earliest ***intime***)
Included (n=53,150)
Exclude (n=100) entries
admitted or transferred to the ICU ***after 2020***
MIMIC-IV Cohort
(n=53,050)

Figure 24: Cohort selection diagram - MIMIC-IV

Table 8: MIMIC-IV cohort characteristics, with count (%) or median (Q1–Q3).

Characteristic		Missingness	Type
Gender			
Female	23,313 (43.9%)	–	categorical
Male	29,737 (56.1%)	–	categorical
Age at Admission	66 (54-78)	0.0%	continuous
O2 Delivery Device(s)			
Use device	33,359 (62.9%)	–	categorical
None	18,549 (35.0%)	–	categorical
Missing	1,142 (2.2%)	–	categorical
Pupil Response R			
Brisk	39,708 (74.9%)	–	categorical
Sluggish	4,603 (8.7%)	–	categorical
Non-reactive	1,812 (3.4%)	–	categorical
Missing	6,927 (13.1%)	–	categorical
first_careunit			
Medical Intensive Care Unit (MICU)	10,213 (19.3%)	–	categorical
Surgical Intensive Care Unit (SICU)	8,241 (15.5%)	–	categorical
Medical/Surgical Intensive Care Unit (MICU/S...	8,808 (16.6%)	–	categorical
Cardiac Vascular Intensive Care Unit (CVICU)	9,437 (17.8%)	–	categorical
Coronary Care Unit (CCU)	6,098 (11.5%)	–	categorical
Trauma SICU (TSICU)	6,947 (13.1%)	–	categorical
Other	3,306 (6.2%)	–	categorical
Anion Gap	13 (11-16)	0.5%	continuous
Heart Rhythm			
SR (Sinus Rhythm)	34,004 (64.1%)	–	categorical
Abnormal heart rhythm	18,657 (35.2%)	–	categorical
Missing	389 (0.7%)	–	categorical
Glucose FS (range 70 -100)	131 (110-164)	32.7%	continuous
Eye Opening			
Spontaneously	39,216 (73.9%)	–	categorical
To Speech	7,387 (13.9%)	–	categorical
None	4,538 (8.6%)	–	categorical
To Pain	1,702 (3.2%)	–	categorical
Missing	207 (0.4%)	–	categorical
Lactate	2 (1-2)	22.0%	continuous
Motor Response			
Obeys Commands	44,409 (83.7%)	–	categorical
Localizes Pain	3,419 (6.4%)	–	categorical
Flex-withdraws	1,673 (3.2%)	–	categorical
No response	2,930 (5.5%)	–	categorical
Abnormal extension	157 (0.3%)	–	categorical
Abnormal Flexion	238 (0.4%)	–	categorical
Missing	224 (0.4%)	–	categorical
Respiratory Pattern			
Regular	29,373 (55.4%)	–	categorical
Not regular	1,739 (3.3%)	–	categorical
Missing	21,938 (41.4%)	–	categorical
Richmond-RAS Scale	0 (-1-0)	15.4%	categorical
in-icu mortality			
0	49,716 (93.7%)	–	categorical
1	3,334 (6.3%)	–	categorical

F.2. Features

18 Gauge Dressing Occlusive
18 Gauge placed in outside facility
20 Gauge Dressing Occlusive
20 Gauge placed in outside facility
20 Gauge placed in the field
Abdominal Assessment
Activity
Activity Tolerance
Admission Weight (Kg)
Admission Weight (lbs.)
Alanine Aminotransferase (ALT)
Alarms On
Albumin
Alkaline Phosphatase
All Medications Tolerated
Ambulatory aid
Anion Gap
Anion gap
Anti Embolic Device
Anti Embolic Device Status
Asparate Aminotransferase (AST)
Assistance
BUN
Balance
Base Excess
Basophils
Bath
Bicarbonate
Bilirubin, Total
Bowel Sounds
Braden Activity
Braden Friction/Shear
Braden Mobility
Braden Moisture
Braden Nutrition
Braden Sensory Perception
CAM-ICU MS Change
Calcium non-ionized
Calcium, Total
Calculated Total CO2
Capillary Refill L
Capillary Refill R
Chloride
Chloride (serum)
Commands
Commands Response
Cough Effort
Cough Type
Creatinine
Creatinine (serum)
Currently experiencing pain
Daily Wake Up
Delirium assessment
Dialysis patient
Diet Type
Difficulty swallowing
Dorsal PedPulse L
Dorsal PedPulse R
ETOH
Ectopy Type 1
Edema Amount
Edema Location
Education Barrier
Education Existing Knowledge
Education Learner
Education Method
Education Readiness/Motivation
Education Response
Education Topic
Eosinophils
Epithelial Cells
Eye Opening
Family Communication
Flatus
GU Catheter Size
Gait/Transferring
Glucose (serum)
Glucose FS (range 70 -100)
Goal Richmond-RAS Scale
HCO3 (serum)
HOB
HR
HR Alarm - High
HR Alarm - Low
Heart Rhythm
Height
Height (cm)
Hematocrit
Hematocrit (serum)
Hemoglobin
History of falling (within 3 mnths)*
History of slips / falls
Home TF
INR
INR(PT)
IV/Saline lock
Insulin pump
Intravenous / IV access prior to admission
Judgement
LLE Color
LLE Temp
LLL Lung Sounds
LUE Color
LUE Temp
LUL Lung Sounds
Lactate
Lactic Acid
Living situation
Lymphocytes

MCH
MCHC
MCV
Magnesium
Mental status
Monocytes
Motor Response
NBP Alarm - High
NBP Alarm - Low
NBP Alarm Source
NBPd
NBPm
NBPs
Nares L
Nares R
Neutrophils
O2 Delivery Device(s)
Oral Care
Oral Cavity
Orientation
PT
PTT
Pain Assessment Method
Pain Cause
Pain Level
Pain Level Acceptable
Pain Level Response
Pain Location
Pain Management
Pain Present
Pain Type
Parameters Checked
Phosphate
Phosphorous
Platelet Count
Position
PostTib Pulses L
PostTib Pulses R
Potassium
Potassium (serum)
Potassium, Whole Blood
Pressure Reducing Device
Pressure Ulcer Present
Pupil Response L
Pupil Response R
Pupil Size Left
Pupil Size Right
RBC
RDW
RLE Color
RLE Temp
RLL Lung Sounds
RR
RUE Color
RUE Temp
RUL Lung Sounds
Radial Pulse L
Radial Pulse R
Red Blood Cells
Resp Alarm - High
Resp Alarm - Low
Respiratory Effort
Respiratory Pattern
Richmond-RAS Scale
ST Segment Monitoring On
Safety Measures
Secondary diagnosis
Self ADL
Side Rails
Skin Color
Skin Condition
Skin Integrity
Skin Temp
Sodium
Sodium (serum)
SpO2
SpO2 Alarm - High
SpO2 Alarm - Low
SpO2 Desat Limit
Specific Gravity
Specimen Type
Speech
Strength L Arm
Strength L Leg
Strength R Arm
Strength R Leg
Support Systems
Temp Site
Temperature F
Therapeutic Bed
Tobacco Use History
Turn
Untoward Effect
Urea Nitrogen
Urine Source
Verbal Response
Visual / hearing deficit
WBC
White Blood Cells
Yeast
admit_age
gender
pCO2
pH
pO2

F.3. Missingness heatmaps

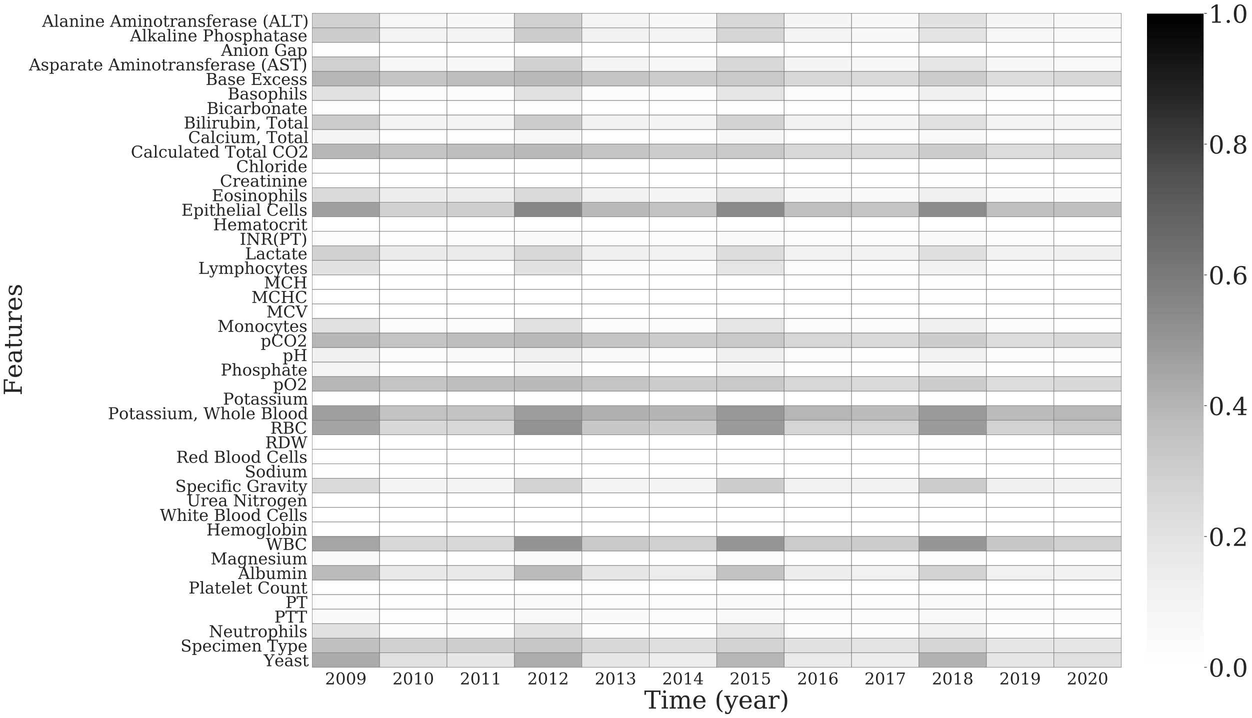


Figure 25: Missingness over time for labevents features in MIMIC-IV dataset after cohort selection. The darker the color, the larger the proportion of missing data.

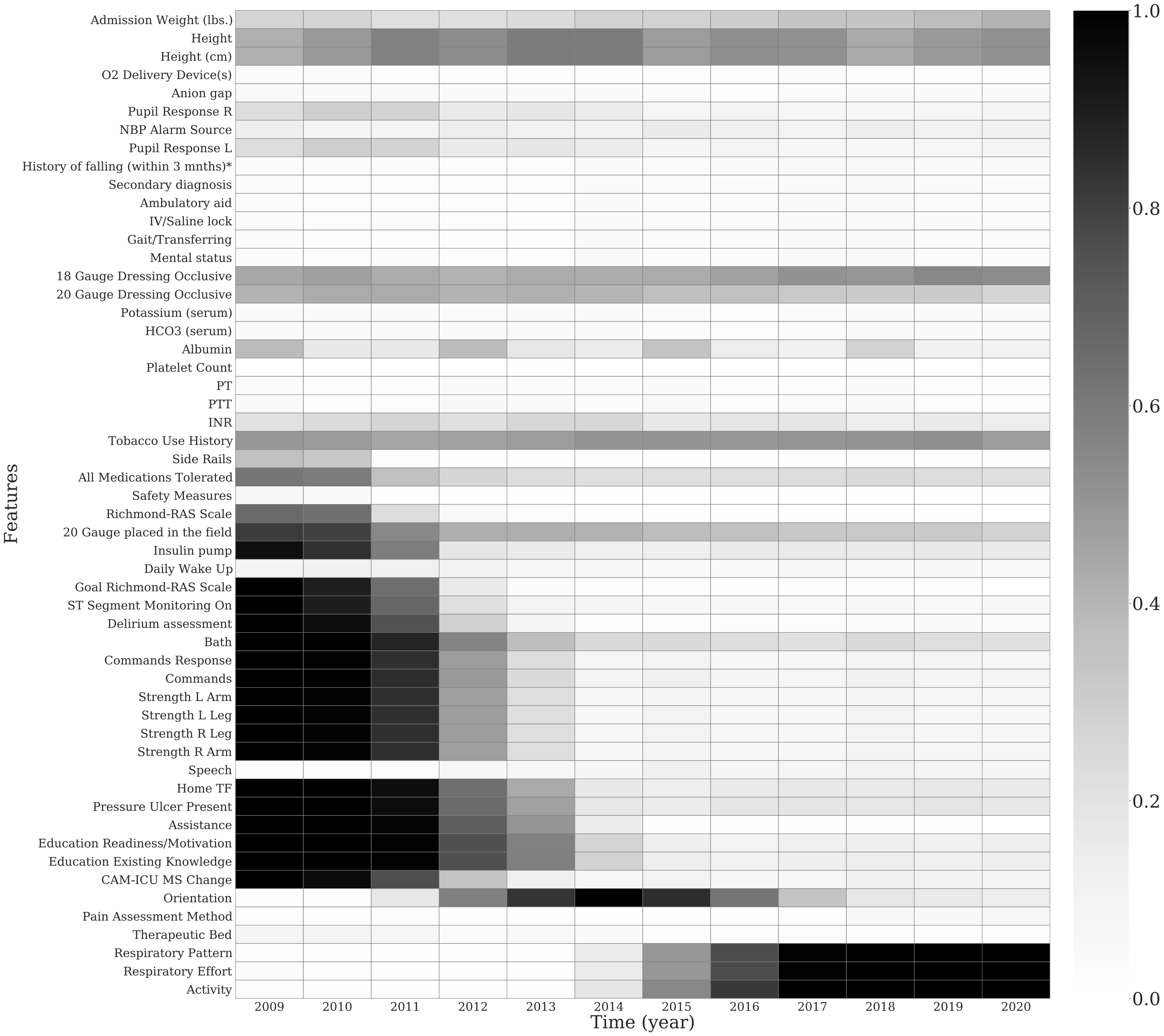


Figure 26: Missingness over time for chartevents features in MIMIC-IV dataset after cohort selection. The darker the color, the larger the proportion of missing data. (part 1)

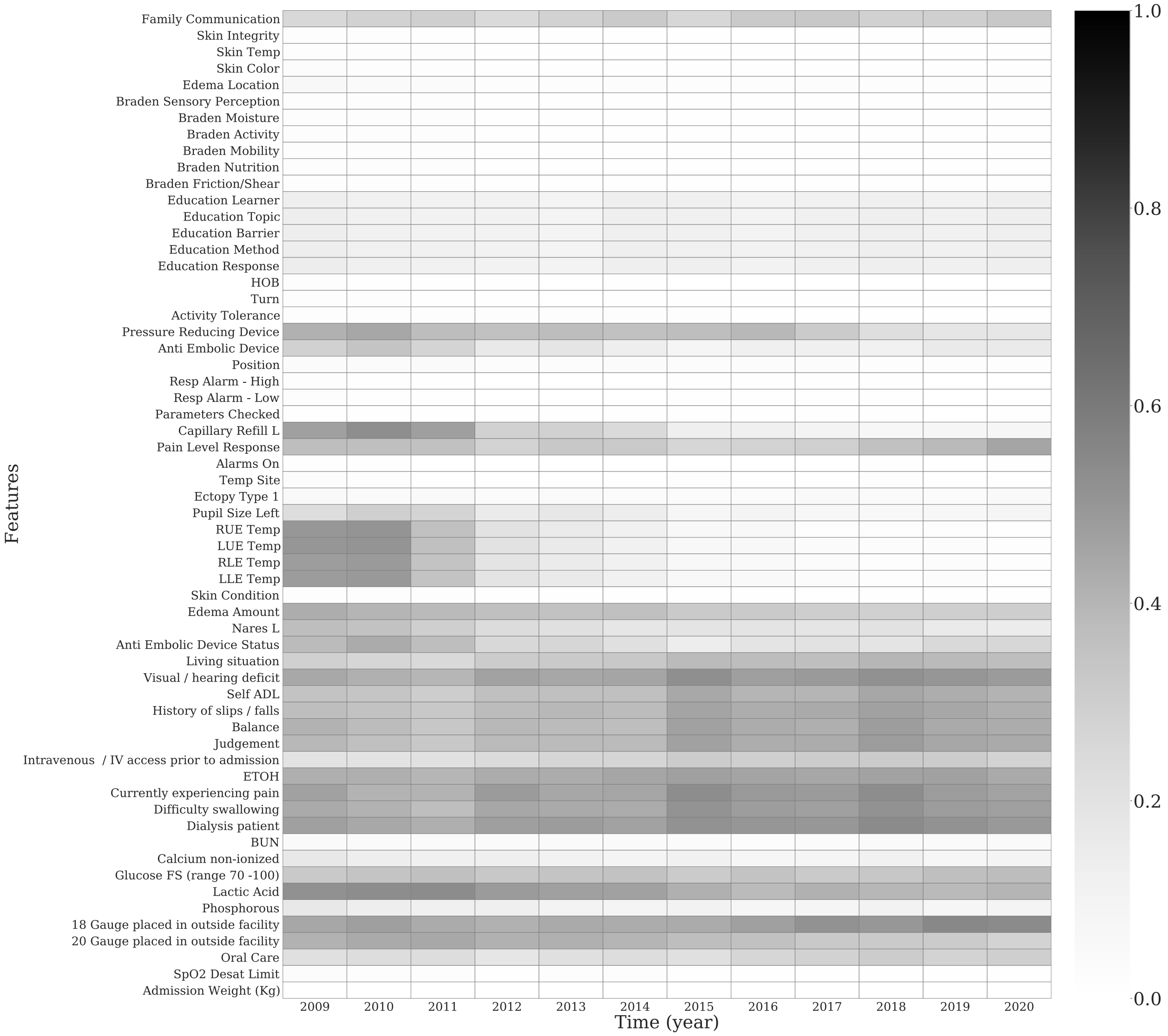


Figure 27: Missingness over time for chartevents features in MIMIC-IV dataset after cohort selection. The darker the color, the larger the proportion of missing data. (part 2)

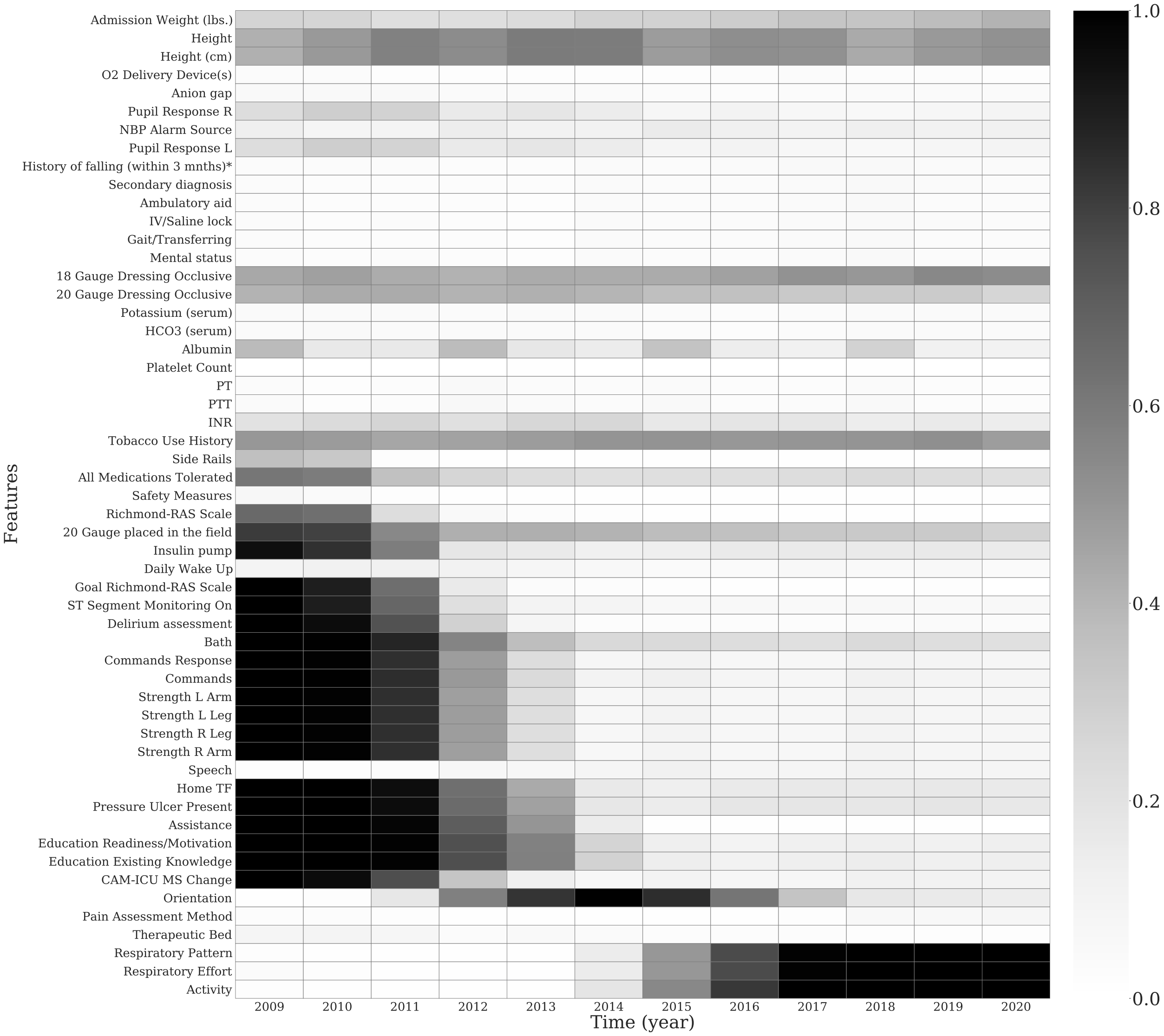


Figure 28: Missingness over time for chartevents features in MIMIC-IV dataset after cohort selection. The darker the color, the larger the proportion of missing data. (part 3)

Appendix G. Additional OPTN (Liver) Data Details

The Organ Procurement and Transplantation Network (OPTN) database Organ Procurement and Transplantation Network (2020) tracks organ donation and transplant events in the U.S. Our study uses data from candidates on the liver transplant wait list. The performance over time is evaluated on a *yearly* basis.

- First, we provide the disclaimer: "The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and transplantation Network. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government".
- Data access: After signing the Data Use Agreement - I from Organ Procedurement And Transplantation network, users can access the OPTN (Liver) dataset.
- Cohort selection: The cohort consists of liver transplant candidates on the waiting list (2005-2017). We follow the same pipeline as Byrd et al. (2021) to extract the data, except that we select the first record for each patient. Cohort selection diagrams are given in Figures 29. This corresponds to a particular interpretation of the prediction: when a patient is first added to the transplant list, given what we know about that patient, what is their estimated risk of 180-day mortality?
- Outcome definition: 180-day mortality from when the patient was first added to the list
- Cohort characteristics: Cohort characteristics are given in Table 9.
- Features: We list the features used in the OPTN liver dataset in Section G.2. We convert all categorical variables into dummy features, and apply standard scaling to numerical variables (subtract mean and divide by standard deviation).
- Missingness heat maps: are given in Figures 30 and 31.

G.1. Cohort Selection and Cohort Characteristics

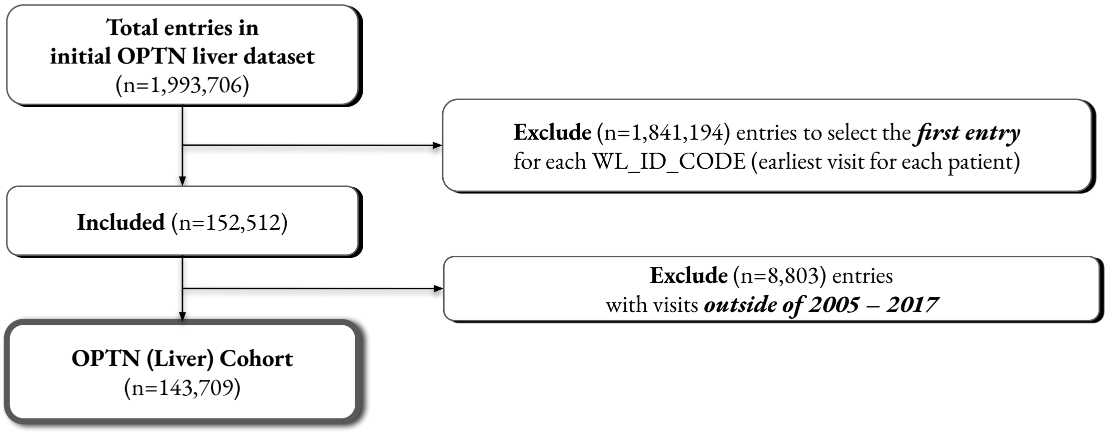

Total entries in
initial OPTN liver dataset
(n=1,993,706)
Exclude (n=1,841,194) entries to select the first entry
for each WL_ID_CODE (earliest visit for each patient)
Included (n=152,512)
Exclude (n=8,803) entries
with visits outside of 2005 – 2017
OPTN (Liver) Cohort
(n=143,709)

Figure 29: Cohort selection diagram - OPTN (Liver)

Table 9: OPTN (Liver) cohort characteristics, with count (%) or median (Q1 – Q3).

Feature name (value)		Empty (ratio)	Type
Gender			
Male	92,560 (64.4%)	–	categorical
Female	51,149 (35.6%)	–	categorical
INIT_AGE	56 (49-62)	0.0%	continuous
FUNC_STAT_TCR	2,070 (2,050-2,080)	0.0%	categorical
INIT_OPO_CTR_CODE	11,036 (3,782-19,282)	0.0%	categorical
ALBUMIN	3 (3-4)	0.0%	continuous
HCC_DIAGNOSIS_TCR			
No	31,390 (21.8%)	–	categorical
Yes	11,312 (7.9%)	–	categorical
Missing	101,007 (70.3%)	–	categorical
PERM_STATE			
CA	19,645 (13.7%)	–	categorical
TX	14,692 (10.2%)	–	categorical
NY	9,976 (6.9%)	–	categorical
GA	4,052 (2.8%)	–	categorical
MD	4,050 (2.8%)	–	categorical
FL	7,602 (5.3%)	–	categorical
PA	8,013 (5.6%)	–	categorical
MI	3,989 (2.8%)	–	categorical
Other	71,007 (49.4%)	–	categorical
EDUCATION	4 (3-5)	0.0%	categorical
ASCITES	2 (1-2)	0.0%	categorical
MORTALITY_180D			
1	4,635 (3.2%)	–	categorical
0	139,074 (96.8%)	–	categorical

G.2. Features

```
ABO
BACT_PERIT_TCR
CITIZENSHIP
DGN_TCR
DGN2_TCR
DIAB
EDUCATION
FUNC_STAT_TCR
GENDER
LIFE_SUP_TCR
MALIG_TCR
OTH_LIFE_SUP_TCR
PERM_STATE
PORTAL_VEIN_TCR
PREV_AB_SURG_TCR
PRI_PAYMENT_TCR
REGION
TIPSS_TCR
VENTILATOR_TCR
WORK_INCOME_TCR
ETHCAT
HCC_DIAGNOSIS_TCR
MUSCLE_WAST_TCR
INIT_OPO_CTR_CODE
WLHR
WLIN
WLKI
WLLU
WLPA
INACTIVE
ASCITES
ENCEPH
DIALYSIS_PRIOR_WEEK
INIT_HGT_CM
INIT_WGT_KG
INIT_BMI_CALC
INIT_AGE
UNOS_CAND_STAT_CD
BILIRUBIN
SERUM_CREAT
INR
SERUM_SODIUM
ALBUMIN
BILIRUBIN_DELTA
SERUM_CREAT_DELTA
INR_DELTA
SERUM_SODIUM_DELTA
ALBUMIN_DELTA
```

G.3. Missingness heatmaps

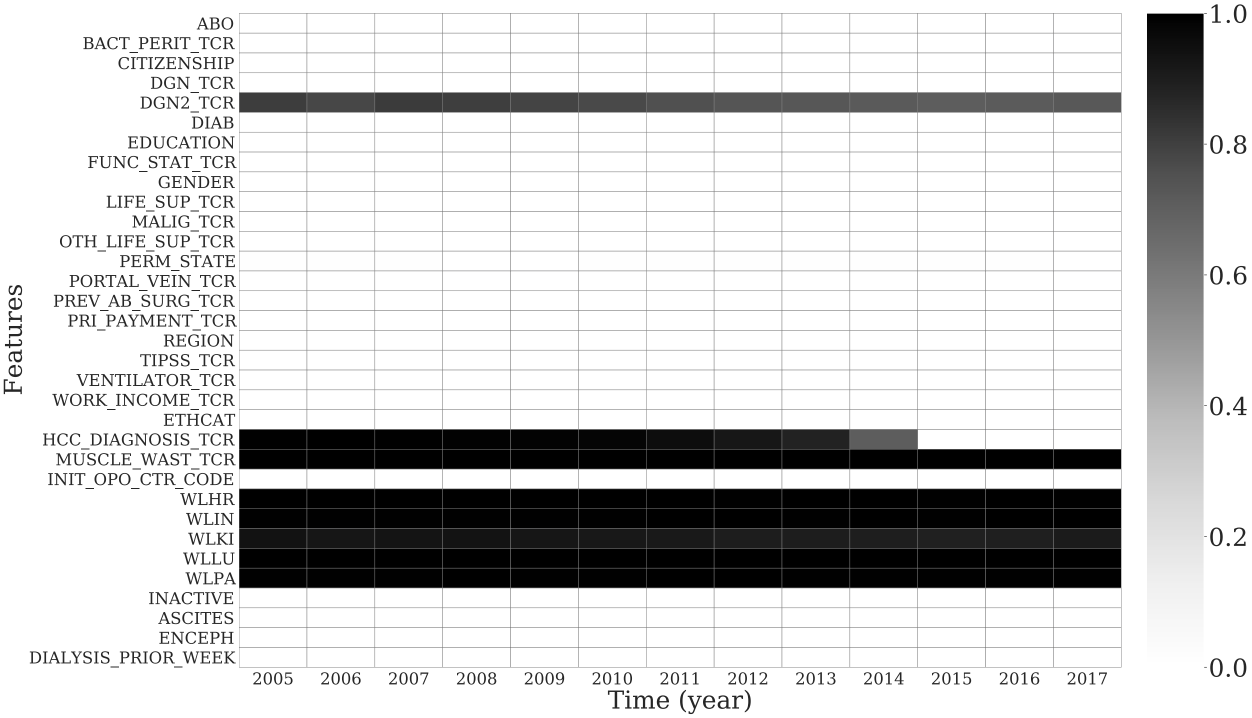


Figure 30: Missingness over time for categorical features in OPTN (Liver) dataset after cohort selection. The darker the color, the larger the proportion of missing data.

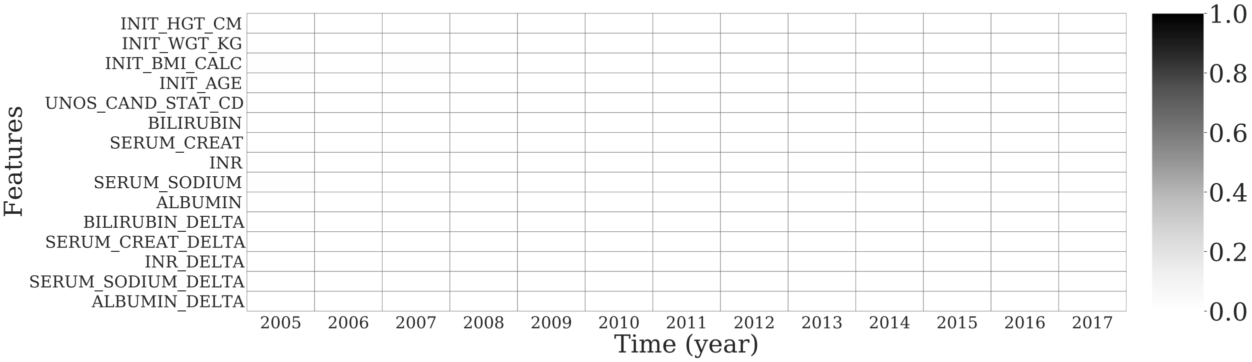


Figure 31: Missingness over time for numerical features in OPTN (Liver) dataset after cohort selection. The darker the color, the larger the proportion of missing data. (Near-zero missingness here.)

Appendix H. Additional MIMIC-CXR Data Details

The MIMIC Chest X-ray (MIMIC-CXR-JPG) (Johnson et al., 2019b) is a publicly available dataset containing chest radiographs in JPG format from 2009–2018. Similar to MIMIC-IV, MIMIC-CXR add time annotations placing each sample into a three-year time range. We approximate the year of each sample by taking the midpoint of its time range. Each patient has an `anchor_year_group`, `anchor_year` and `StudyDate`. For each patient, we first calculated an offset as the difference between `StudyDate` and `anchor_year`. Then, we approximated the admit time as the midpoint of `anchor_year_group` after applying the computed offset. The performance over time is evaluated on a *yearly* basis. Our study uses MIMIC-IV-JPG-2.0. A similar training setup to that in Seyyed-Kalantari et al. (2020) was used (learning rate, architecture, data augmentation, stopping criteria, etc.).

- Data access: Users must create a Physionet account, become credentialed, and sign a data use agreement (DUA).
- Cohort selection: We removed the records from 2009 due to the tiny sample size. (Selection diagram in Figure 32). We keep all records for each patients and split the data based on patient `subject id`.
- Outcome definition: The outcome is the probabilities of all labels given the input images. The labels includes 13 abnormal outcomes and 1 normal outcome. (Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices, No Finding)
- Cohort characteristics: Cohort characteristics are given in Table 10.

H.1. Cohort Selection and Cohort Characteristics

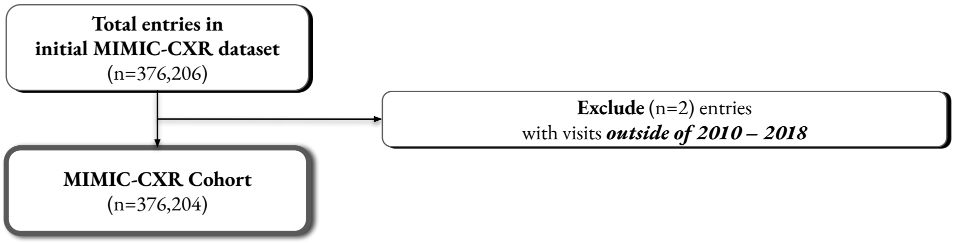


Figure 32: Cohort selection diagram - MIMIC-CXR

Table 10: MIMIC-CXR cohort characteristics, with count (%) or median (Q1–Q3).

Feature name (value)	Summary statistic	Empty (ratio)	Status
Gender			
F	179,765 (47.8%)	–	categorical
M	196,439 (52.2%)	–	categorical
Age	64 (51-76)	0.0%	continuous
Diseases			
Atelectasis	65,390 (17.4%)	–	categorical
Cardiomegaly	56,404 (15.0%)	–	categorical
Consolidation	14,394 (3.8%)	–	categorical
Edema	36,026 (9.6%)	–	categorical
Enlarged Cardiomediastinum	9,821 (2.6%)	–	categorical
Fracture	6,314 (1.7%)	–	categorical
Lung Lesion	10,574 (2.8%)	–	categorical
Lung Opacity	76,074 (20.2%)	–	categorical
Pleural Effusion	75,526 (20.1%)	–	categorical
Pleural Other	3,432 (0.9%)	–	categorical
Pneumonia	25,065 (6.7%)	–	categorical
Pneumothorax	12,828 (3.4%)	–	categorical
Support Devices	69,148 (18.4%)	–	categorical
No Finding	167,116 (44.4%)	–	categorical

H.2. Label level AUROC over time for MIMIC-CXR

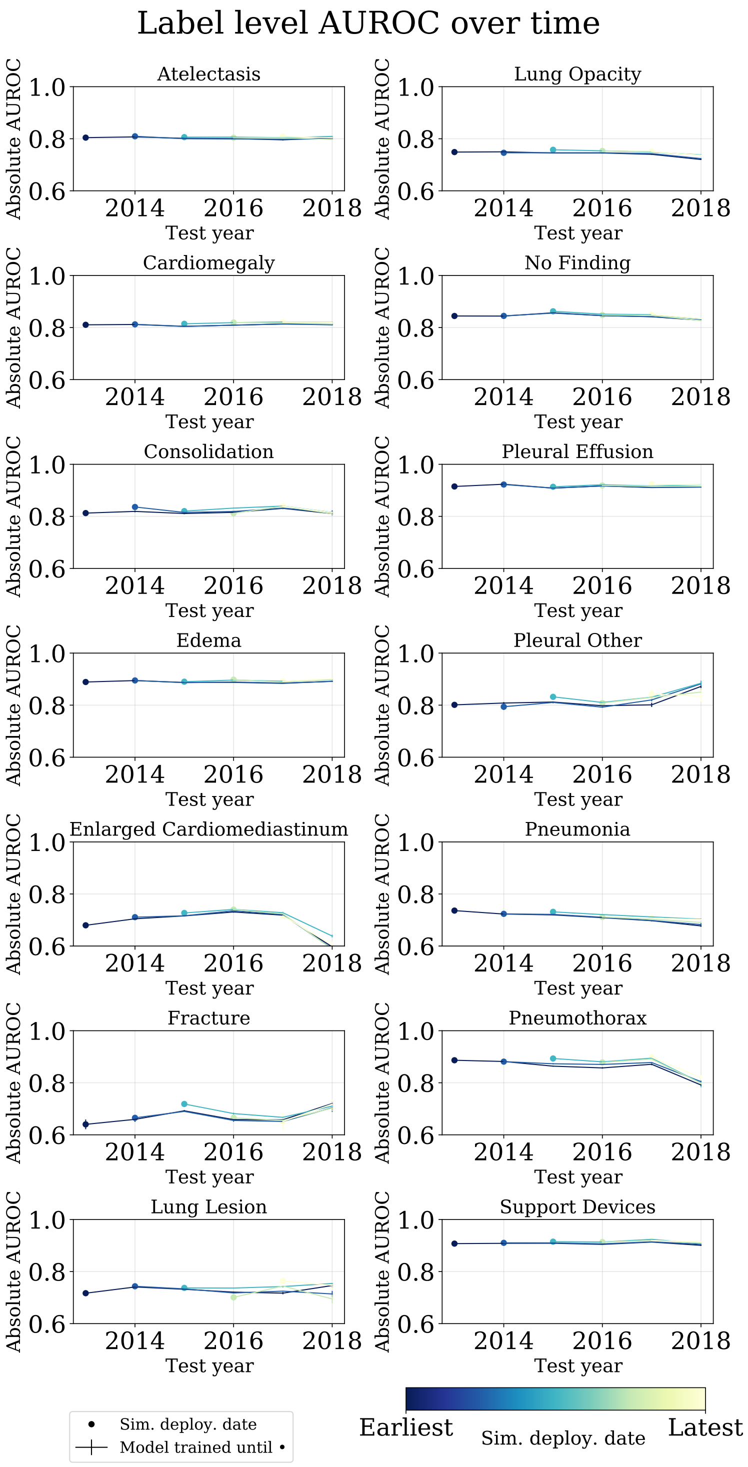


Figure 33: Absolute AUROC over time of each label in MIMIC-CXR

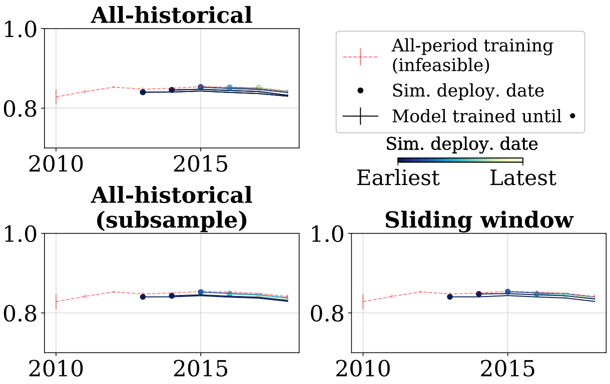


Figure 34: Weighted test AUROC vs. year for the DenseNet architecture on MIMIC-CXR.

Table 11: MIMIC-CXR label-level AUROC from time-agnostic evaluation of all-period training. The format is mean (±std. dev. across splits)

Label	**AUROC**	**Label**	**AUROC**
Atelectasis	0.826 (±0.003)	Cardiomegaly	0.837 (±0.002)
Consolidation	0.841 (±0.003)	Edema	0.904 (±0.002)
Enlarged Cardiomediastinum	0.759 (±0.005)	Fracture	0.745 (±0.006)
Lung Lesion	0.784 (±0.003)	Lung Opacity	0.770 (±0.002)
Pleural Effusion	0.929 (±0.001)	Pleural Other	0.844 (±0.009)
Pneumonia	0.755 (±0.004)	Pneumothorax	0.918 (±0.006)
Support Devices	0.928 (±0.001)	No Finding	0.876 (±0.002)

Appendix I. Logistic Regression Coefficients from Splitting by Patient

To help with intuition in important features for the predictive task on each dataset, here we have the coefficients of logistic regression models trained from splitting by patient.

Table 12: SEER (Breast) top 10 important features for LR models, all-period training.

Feature	Coefficient
SEER historic stage A (1973-2015)_Distant	-2.113944
SEER historic stage A (1973-2015)_Localized	1.676493
Regional nodes examined (1988+)_95.0	-1.167844
CS lymph nodes (2004-2015)_750	1.100824
CS lymph nodes (2004-2015)_755	1.023753
Histologic Type ICD-O-3_8530	-0.913494
Histologic Type ICD-O-3_8543	0.902798
Breast - Adjusted AJCC 6th T (1988-2015)_T4d	0.899491
Histologic Type ICD-O-3_8211	0.877848
EOD 10 - extent (1988-2003)_85	-0.791136

Table 13: SEER (Colon) top 10 important features for LR models, all-period training.

Feature	Coefficient
Reason no cancer-directed surgery_Surgery performed	2.360161
Regional nodes positive (1988+)_00	1.897706
Regional nodes positive (1988+)_01	1.872008
modified AJCC stage 3rd (1988-2003)_40	-1.787481
EOD 10 - extent (1988-2003)_13	1.766066
Reason no cancer-directed surgery_Not recommended, contraindicated due to other cond; autopsy only (1973-2002)	-1.752474
EOD 10 - extent (1988-2003)_85	-1.732619
EOD 10 - extent (1988-2003)_70	-1.704333
CS mets at dx (2004-2015)_99	1.619905
CS mets at dx (2004-2015)_00	1.609454

Table 14: SEER (Lung) top 10 important features for LR models, all-period training.

Feature	Coefficient
Histologic Type ICD-O-3_8240	2.514539
EOD 4 - nodes (1983-1987)_0	2.074730
EOD 4 - nodes (1983-1987)_7	-1.777530
EOD 10 - size (1988-2003)_140	-1.587893
Histologic Type ICD-O-3_8141	-1.546566
CS tumor size (2004-2015)_998.0	-1.515856
EOD 4 - nodes (1983-1987)_6	-1.497022
Type of Reporting Source_Nursing/convalescent home/hospice	-1.338998
CS mets at dx (2004-2015)_51	-1.326595
EOD 10 - size (1988-2003)_150	-1.326196

Table 15: CDC COVID-19 top 10 important features for LR models, all-period training.

Feature	Coefficient
res_state_DE	2.202055
age_group_0 - 9 Years	-2.114818
age_group_80+ Years	1.965279
age_group_10 - 19 Years	-1.681099
res_state_GA	1.391469
age_group_70 - 79 Years	1.379589
res_county_WICHITA	1.290644
age_group_20 - 29 Years	-1.189734
res_county_SUMNER	-1.135073
mechvent_yn_Yes	1.117372

Table 16: SWPA COVID-19 top 10 important features for LR models according to experiments splitting by patient.

Feature	Coefficient
age_bin_(70, 200]_0	-0.781337
age_bin_(70, 200]_1	0.780673
medication_FENTANYL (PF) 50 MCG/ML INJECTION SOLUTION_0.0	0.651419
medication_EPINEPHRINE 0.3 MG/0.3 ML INJECTION, AUTO-INJECTOR_nan	-0.627565
medication_HYDROCORTISONE SOD SUCCINATE (PF) 100 MG/2 ML SOLUTION FOR INJECTION_0.0	0.544222
medication_HYDROCODONE 5 MG-ACETAMINOPHEN 325 MG TABLET_nan	-0.520368
medication_DEXAMETHASONE SODIUM PHOSPHATE 4 MG/ML INJECTION SOLUTION_0.0	0.502954
medication_ASPIRIN 81 MG TABLET,DELAYED RELEASE_nan	-0.479100
bmi_nan	-0.427569
age_bin_(60, 70]_0	-0.380688

Table 17: MIMIC-IV top 10 important features for LR models, all-period training.

Feature	Coefficient
O2 Delivery Device(s)_None	-0.307334
Eye Opening_None	0.301737
admit_age	0.299712
O2 Delivery Device(s)_Nasal cannula	-0.248463
Motor Response_Obeys Commands	-0.230931
Pupil Response L_Non-reactive	0.223776
Richmond-RAS Scale_ 0 Alert and calm	-0.205476
Temp Site_Blood	-0.204514
HR_0.0	0.197299
Diet Type_NPO	0.195156

Table 18: OPTN (Liver) top 10 important features for LR models, all-period training.

Feature	Coefficient
SERUM_CREAT_DELTA	0.660589
FUNC_STAT_TCR_2020.0	0.241507
FUNC_STAT_TCR_2080.0	-0.236288
DGNC_4110.0	-0.234680
REGION_5.0	0.223940
EDUCATION_998.0	0.218549
ASCITES_3.0	0.218329
ASCITES_1.0	-0.214076
INIT_OPO_CTR_CODE_1054	-0.209265
INIT_OPO_CTR_CODE_4743	-0.207778

Appendix J. Diagnostic plots

We took the union of the top k most important features from each time point to be included in the diagnostic plots, where k was tuned depending on the dataset so that the resulting plots would not be overcrowded. For categorical features, we additionally highlighted (using a thicker line) features that had consistently high prevalence ($\geq p$) or experienced a large change in prevalence across one time point ($\geq \Delta$). The specific parameters of each dataset are defined in each subsection. For numerical features, we highlighted features whose average ranking across all time points was ≤ 3 (also chosen to avoid overcrowding).

J.1. SEER (Breast)

For SEER (Breast) diagnostic plots, important features were selected using $k = 5, p = 0.4, \Delta = 0.2$.

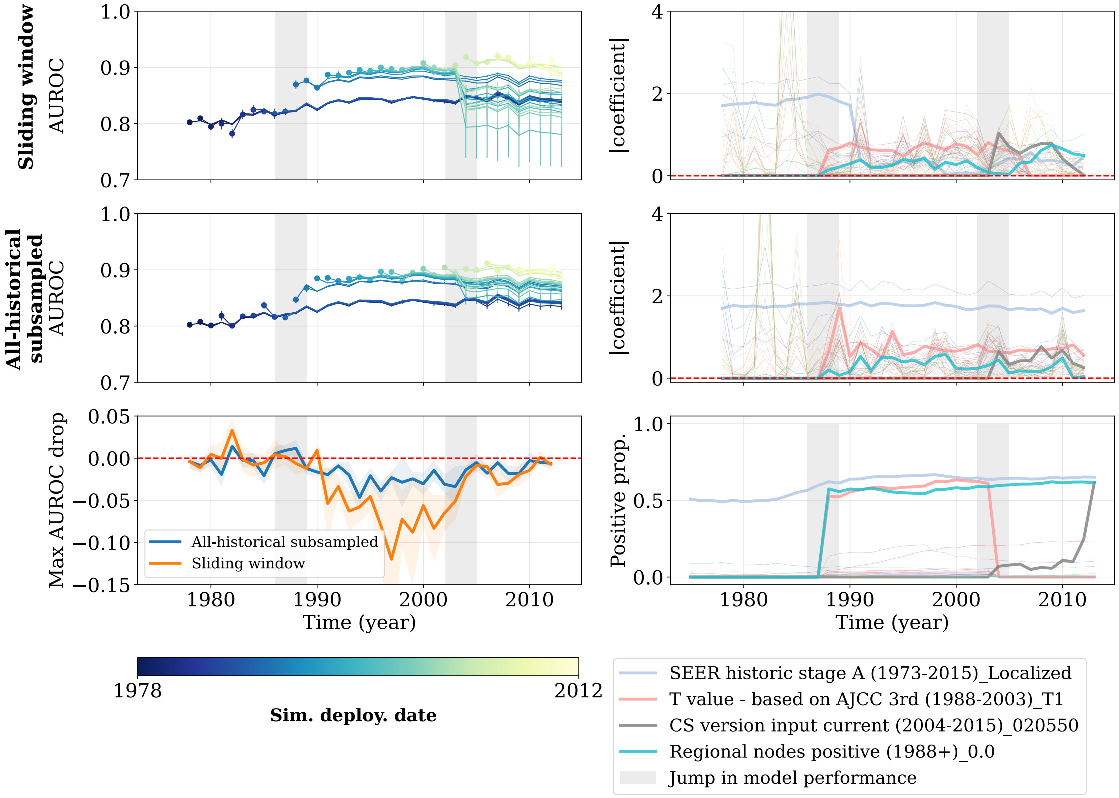


Figure 35: Diagnostic plot of SEER (Breast) dataset. The important features are selected as the union of the top 5 features that have the highest absolute value model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. As shown in the gray highlighted region, there are jumps in performance around 1988 and 2003, which coincides with the introducing and removal of several features (e.g. T value - based on AJCC 3rd (1988-2003)_T1). The latency of jumps in coefficients are caused by length of sliding window.

J.2. SEER (Colon)

For SEER (Colon) diagnostic plots, important features were selected using $k = 3, p = 0.4, \Delta = 0.2$.

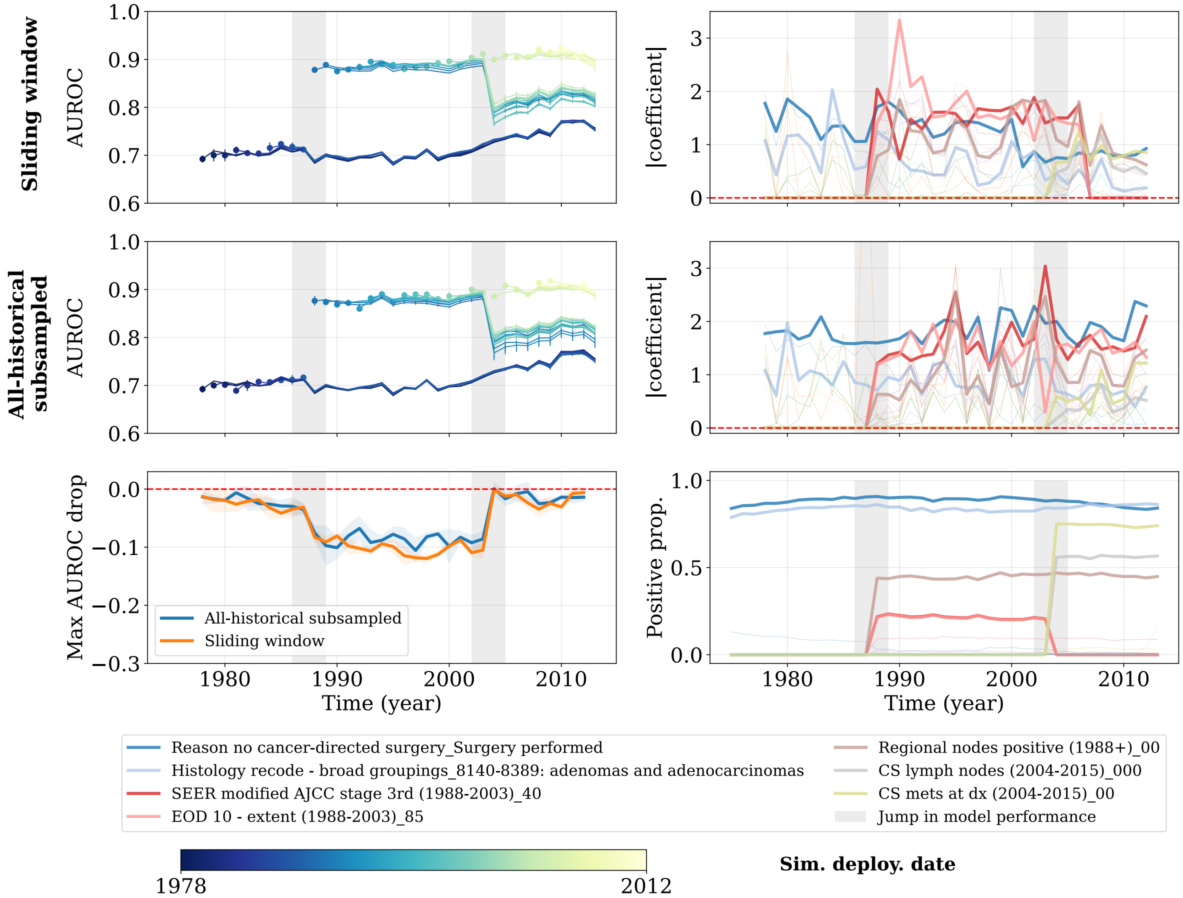


Figure 36: Diagnostic plot of SEER (Colon) dataset. The important features are selected as the union of the top 3 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. As shown in the gray highlighted region, there are jumps in performance around 1988 and 2003, which coincides with the introducing and removal of several features (e.g. SEER modified AJCC stage 3rd (1988-2003)_40). The latency of jumps in coefficients are caused by length of sliding window.

J.3. SEER (Lung)

For SEER (Lung) diagnostic plots, important features were selected using $k = 5, p = 0.2, \Delta = 0.2$.

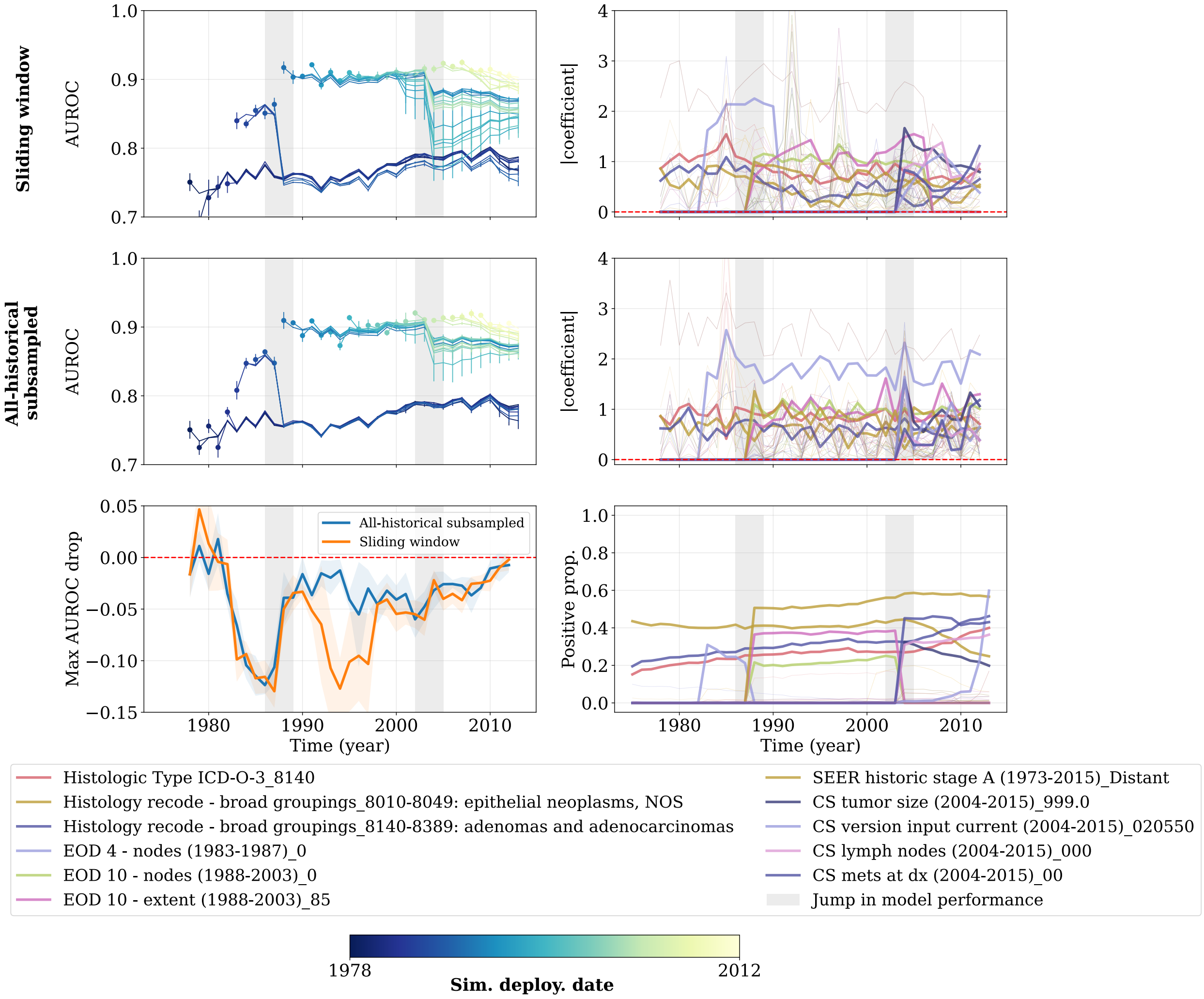


Figure 37: Diagnostic plot of SEER (Lung) dataset. The important features are selected as the union of the top 5 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. As shown in the gray highlighted region, there are jumps in performance around 1988 and 2003, which coincides with the introducing and removal of several features (e.g. EOD 10 - nodes (1988-2013)_0 & EOD 10 - extent (1988-2003)_85). The latency of jumps in coefficients are caused by length of sliding window.

J.4. CDC COVID-19

For CDC COVID-19 diagnostic plots, important features were selected using $k = 5, p = 0.15, \Delta = 0.15$.

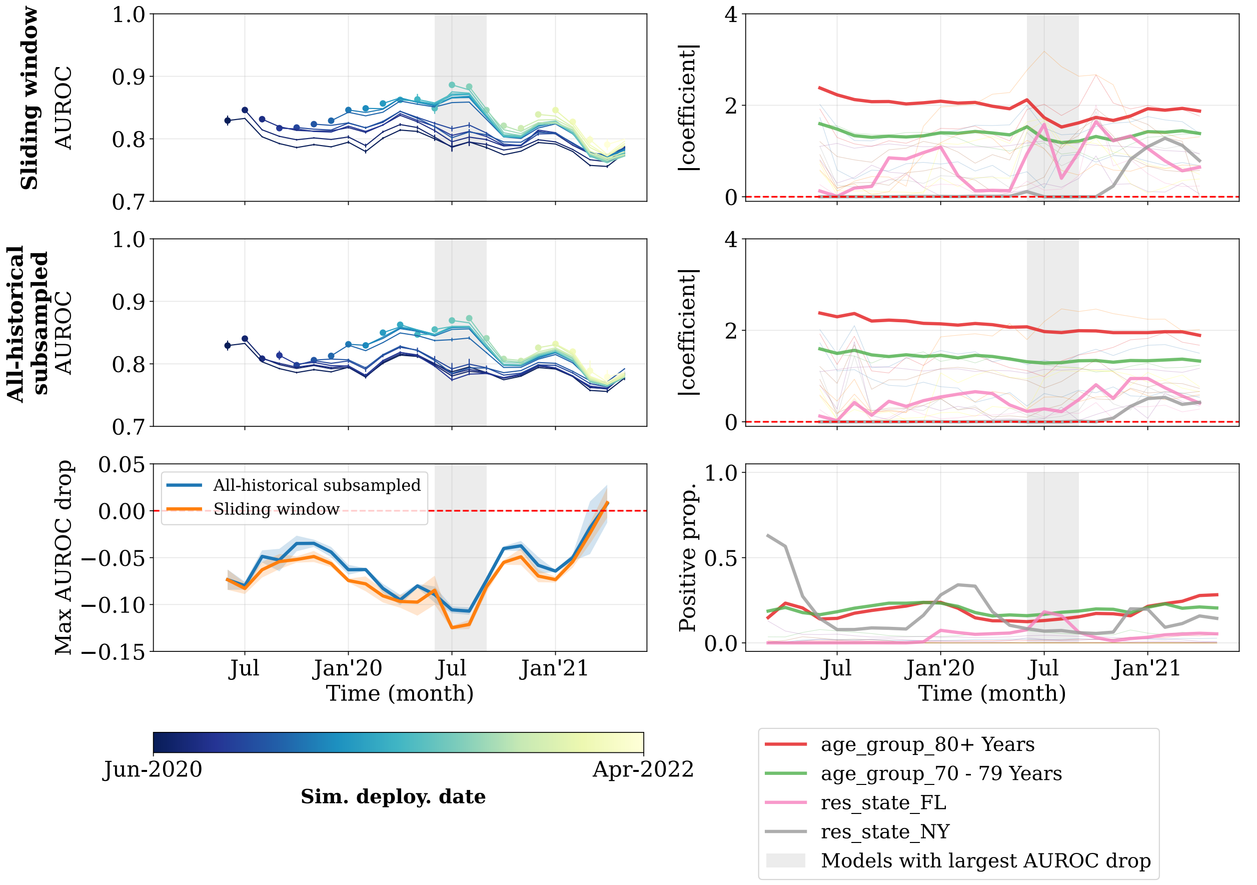


Figure 38: Diagnostic plot of CDC COVID-19. The important features are selected as the union of the top 5 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. As shown in the gray highlighted region, the models trained around June 2021 suffer the largest maximum AUROC drop, coinciding with a shift in distribution of ages (Figure 18(a)) and states (Figure 18(b)). The latency of jumps in coefficients are caused by length of sliding window.

J.5. SWPA COVID-19

For SWPA COVID-19 diagnostic plots, important features were selected using $k = 3, p = 0.4, \Delta = 0.2$.

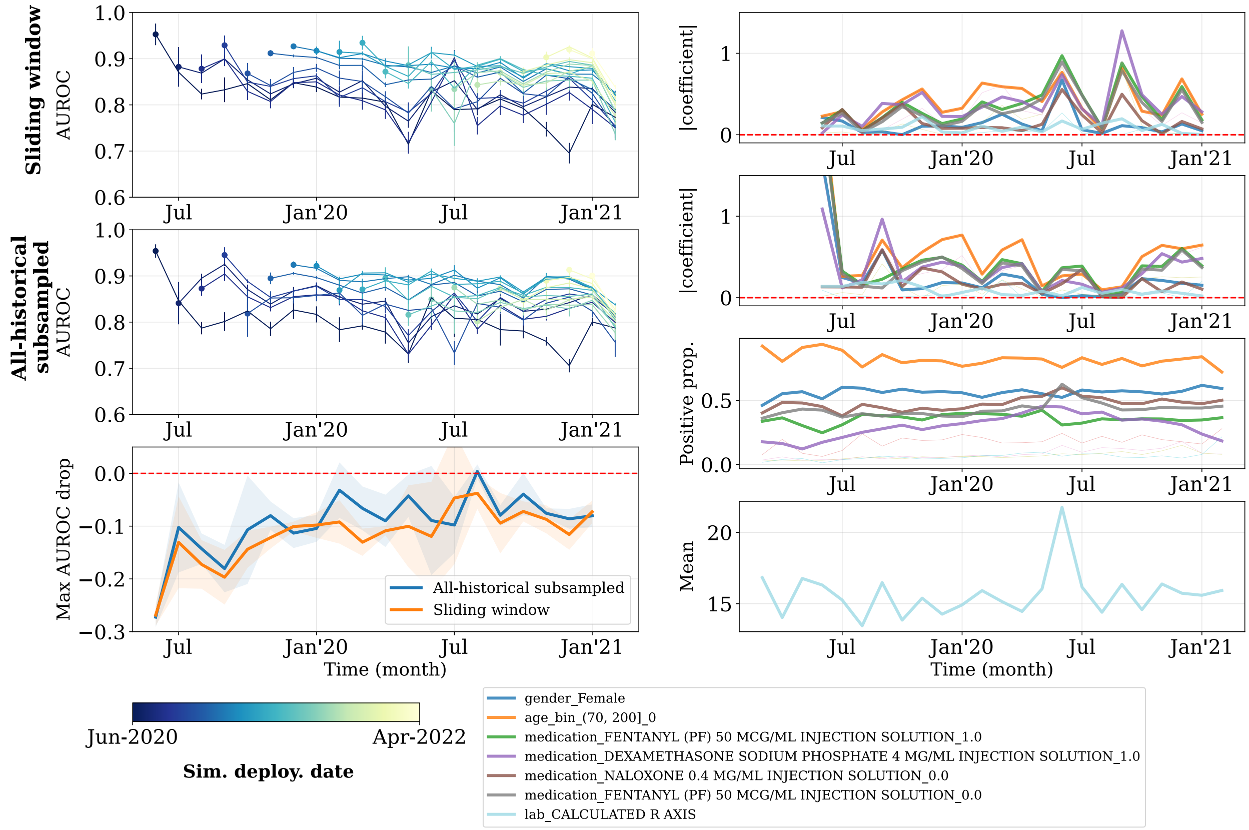


Figure 39: Diagnostic plot of SWPA COVID-19. The important features are selected as the union of the top 3 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. One of the hypotheses for relatively large uncertainty is smaller sample size.

J.6. MIMIC-IV

For MIMIC-IV diagnostic plots, important features were selected using $k = 3, p = 0.4, \Delta = 0.2$.

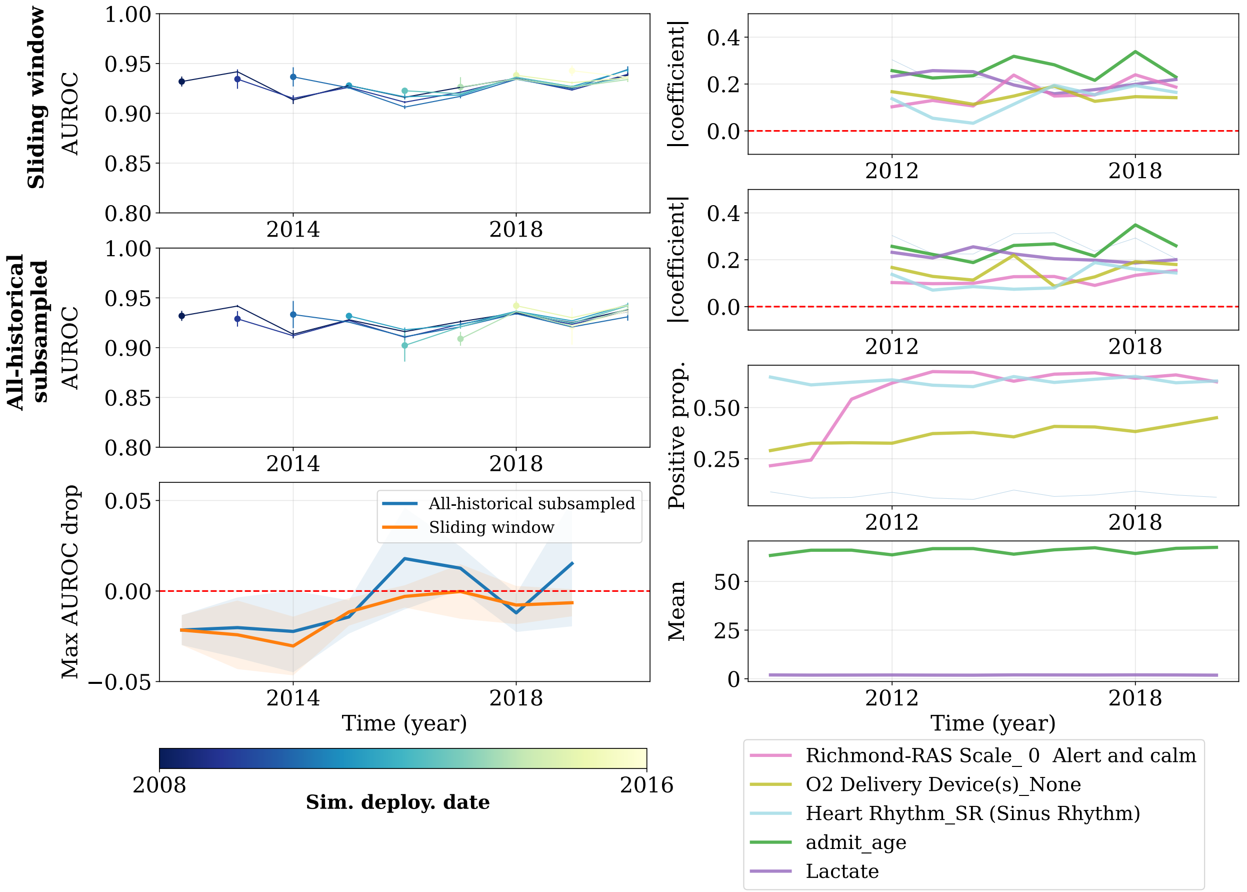


Figure 40: Diagnostic plot of MIMIC-IV. The important features are selected as the union of the top 3 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. The model performance is relatively stable, coinciding with relatively stable distributions of a majority of important features.

J.7. OPTN (Liver)

For OPTN (Liver) diagnostic plots, important features were selected using $k = 3, p = 0.4, \Delta = 0.2$.

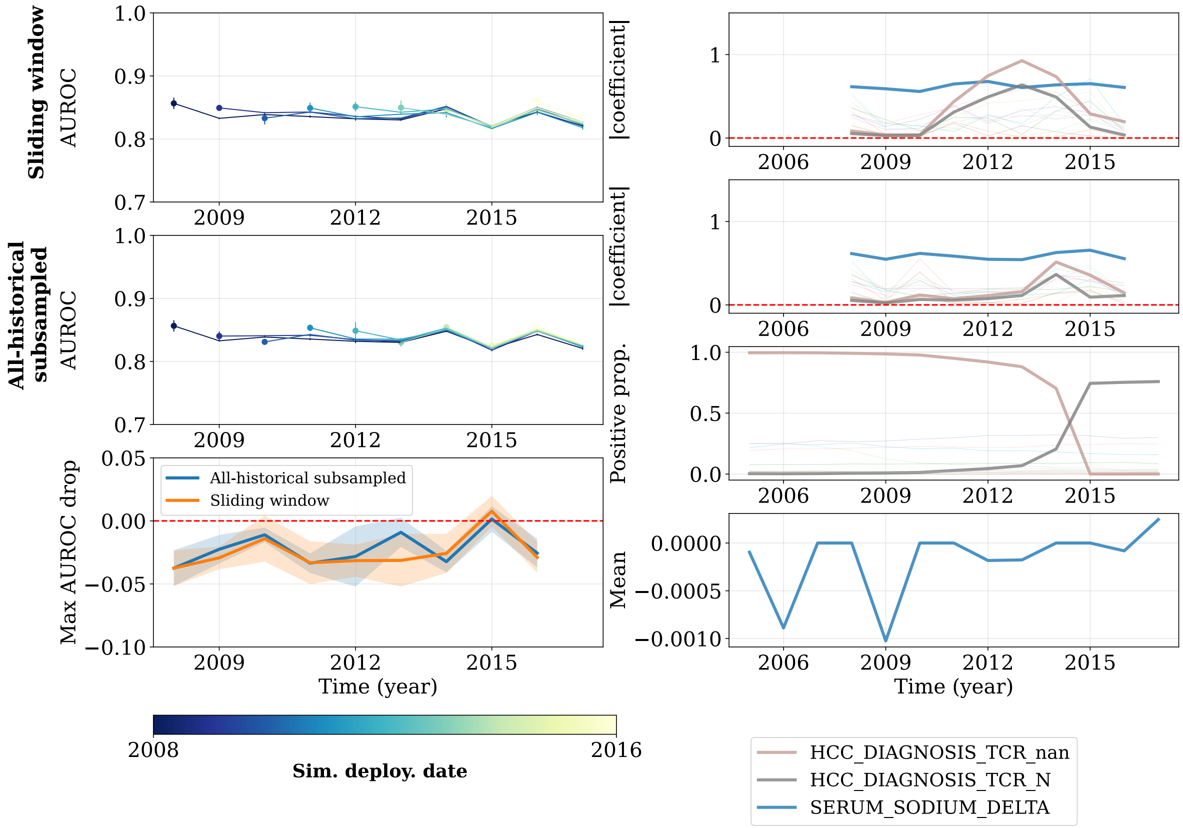


Figure 41: Diagnostic plot of OPTN (Liver). The important features are selected as the union of the top 3 features that have the highest absolute model coefficients. The left column includes AUROC versus time for both sliding window and all-historical subsampled, and the maximum AUROC drop for each trained model. The right column provides the absolute coefficients of each trained model from both regimes, and positive proportion of the significant features over time. Although the HCC DIAGNOSIS TCR binary features change in positive proportion over time, these features were not always important, and the other important features (faded) maintain relatively stable proportions across time. Overall, model performance is quite stable over time.

J.8. MIMIC-CXR

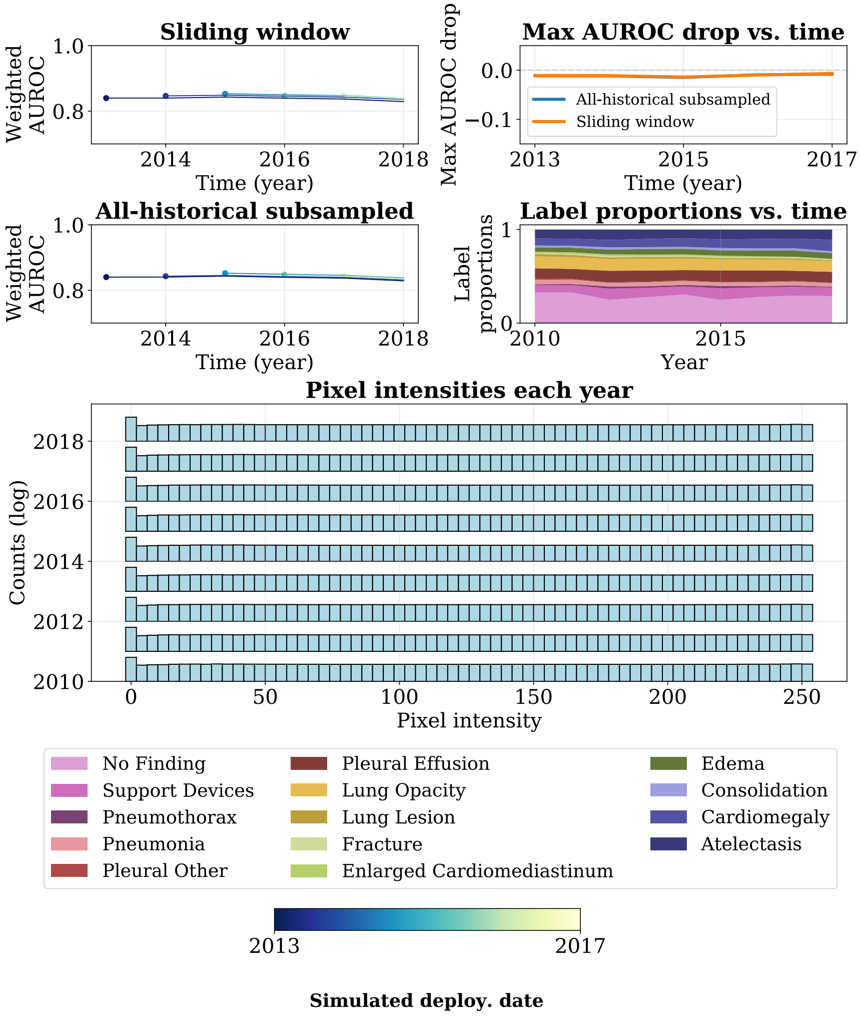


Figure 42: Diagnostic plot of MIMIC-CXR. The top and mid left includes AUROC versus time for both sliding window and all-historical subsampled. The top right is the maximum AUROC drop for each trained model. The mid-right provides the label proportions over time. The bottom shows pixel intensities for images in each year. The histogram of pixel intensity is stable over time, which is consistent with the small variation in model performance over time

Appendix K. Model performance over time from three models

K.1. AUROC

All plots in this section are for the all-historical training regime.

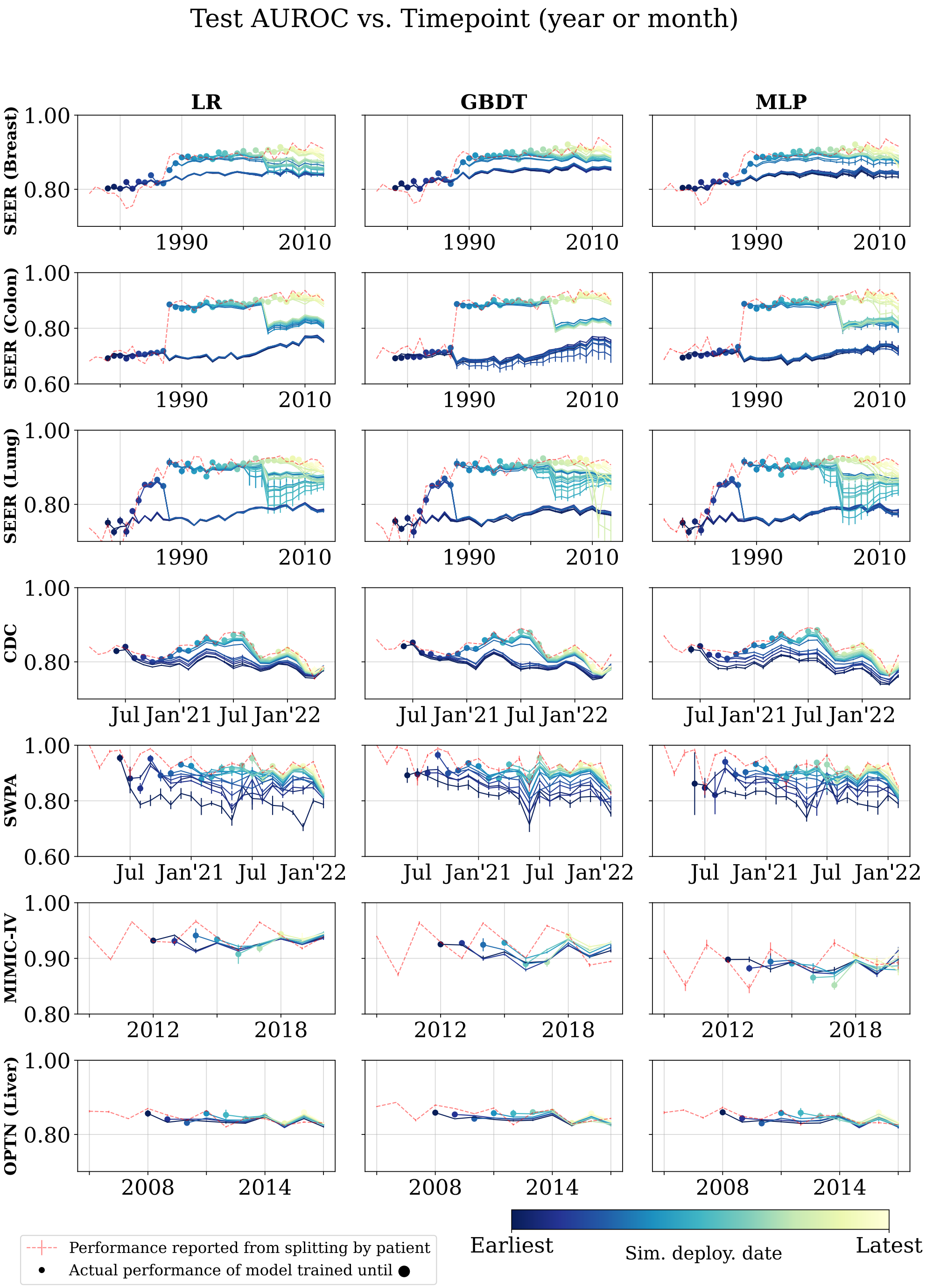


Figure 43: AUROC versus test timepoints from three model classes on all datasets.

K.2. AUPRC

All plots in this section are for the all-historical training regime.

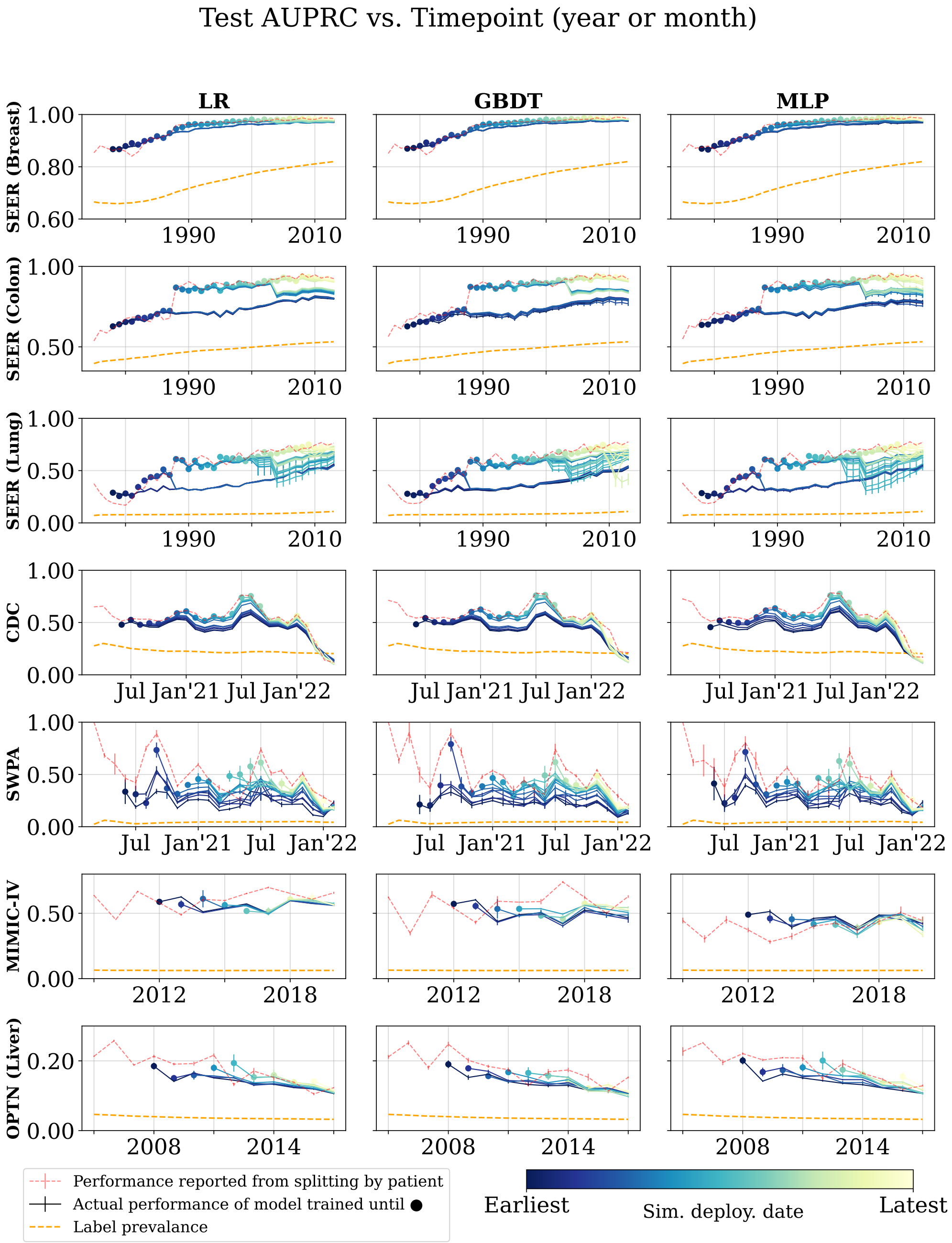


Figure 44: AUPRC versus test timepoints from three model classes on all datasets. Label prevalance refers to the ratio of accumulated positive labels over time.

Appendix L. Data Split Details

Table 19: Split ratio for each dataset for training, validation and testing (both for time-agnostic splits and in-period splits).

Dataset	Split ratio
SEER (Breast)	0.8-0.1-0.1
SEER (Colon)	0.8-0.1-0.1
SEER (Lung)	0.8-0.1-0.1
CDC COVID-19	0.8-0.1-0.1
SWPA COVID-19	0.5-0.25-0.25
MIMIC-IV	0.5-0.25-0.25
OPTN (Liver)	0.5-0.25-0.25
MIMIC-CXR	0.5-0.25-0.25

Appendix M. Hyperparameter Grids

Table 20: Hyperparameter grids for model training.

Parameter	Values Considered
LR	
C	0.01, 0.1, 1, 10, 10^2, 10^3, 10^4, 10^5
GBDT	
n_estimators	50, 100
max_depth	3, 5
learning_rate	0.01, 0.1
MLP	
hidden_layer_sizes	3, 5
learning_rate_init	10^{-4}, 10^{-3}, 0.01

Appendix N. AUROC from full-period training

Table 21: AUROC report from full-period training, the results are in format mean (±std. dev. across splits)

Dataset	Model	Full-period AUROC
SEER (Breast)	LR	0.888 (±0.002)
	GBDT	0.891 (±0.002)
	MLP	0.891 (±0.002)
SEER (Colon)	LR	0.863 (±0.003)
	GBDT	0.868 (±0.002)
	MLP	0.869 (±0.003)
SEER (Lung)	LR	0.894 (±0.002)
	GBDT	0.894 (±0.002)
	MLP	0.898 (±0.002)
CDC COVID-19	LR	0.837 (±0.001)
	GBDT	0.851 (±0.001)
	MLP	0.852 (±0.002)
SWPA COVID-19	LR	0.928 (±0.005)
	GBDT	0.930 (±0.004)
	MLP	0.928 (±0.006)
MIMIC-IV	LR	0.935 (±0.003)
	GBDT	0.931 (±0.002)
	MLP	0.898 (±0.008)
OPTN (Liver)	LR	0.846 (±0.005)
	GBDT	0.854 (±0.005)
	MLP	0.847 (±0.006)
MIMIC-CXR	DenseNet	0.860 (±0.001)